Étude de la variabilité du génome mitochondrial comme facteur de susceptibilité au cancer du sein

Sophie Blein

N° 240-2014 Année 2014

Thèse de l'université de Lyon

Présentée

devant L'UNIVERSITÉ CLAUDE BERNARD LYON1

pour l'obtention

du Diplôme de doctorat
(arrêté du 7 août 2006)

soutenue publiquement le
14 Novembre 2014

par

Sophie Blein

Étude de la variabilité du génome mitochondrial comme facteur de susceptibilité au cancer du sein

Directeur de thèse : David COX

Jury :

Françoise Clavel	Rapporteur
David Cox	Directeur de thèse
Emmanuelle Genin	Rapporteur
Christine Lasset	Présidente du jury
Pascal Reynier	Examinateur

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. François-Noël GILLY**
Vice-président du Conseil d'Administration	M. le Professeur Hamda BEN HADID
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur Philippe LALLE
Vice-président du Conseil Scientifique	M. le Professeur Germain GILLET
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme. la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. FLEURY
Département Chimie Biochimie	Directeur : Mme Caroline FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. Georges TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. Jean-Claude PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. P. FOURNIER
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

REMERCIEMENTS

Ces 3 années de thèse s'achèvent, et celle-ci ne se serait sans doute pas aussi bien déroulée si de nombreuses personnes ne m'avaient pas apporté leur aide, leur soutien, leur présence et leur amitié.

Je tiens à remercier en tout premier lieu mon directeur de thèse, David COX, sans qui je n'aurais tout simplement pas eu l'opportunité de réaliser ma thèse. Merci de t'être battu pour me trouver un financement - ce qui n'a vraiment pas été simple - ainsi que de m'avoir encadrée et formée en épidémiologie génétique. Sous ta tutelle, j'ai eu l'occasion d'interagir avec de nombreux scientifiques internationaux, d'acquérir de l'expérience, et de prendre des responsabilités dans mon travail. Merci pour la grande liberté et l'autonomie que m'a laissées, cela a été extrêmement formateur.

D'autre part, je remercie sincèrement Madame Catherine Lasset et Monsieur Pascal Reynier d'avoir accepté de faire partie de mon jury, ainsi que tout particulièrement Madame Emmanuelle Genin et Madame Françoise Clavel-Chapelon pour avoir examiné mon manuscrit de thèse et m'avoir apporté leurs conseils et leur expertise.

Je remercie également l'ensemble de l'équipe « Génétique du cancer du sein » pour m'avoir accueillie au sein de leur équipe, et m'avoir parfois apporté un point de vue différent en tant que biologistes.

Ces années resteront pour moi associées à une excellente ambiance de travail, et ce notamment grâce aux personnes de la plateforme Synergie Lyon Cancer et du Centre Léon Béard que j'ai côtoyées au quotidien : Emilie T., Janice, Anthony, Anne-Sophie, Jean-Phillipe, Alexia, Sandrine, Vincent et Emilie S, mais aussi Valérie pour sa gentillesse et se disponibilité, et Elise pour son dynamisme et sa bonne humeur permanente. Merci de m'avoir fait partagé votre expertise en bioinformatique et biostatistiques, mais également pour tous les bons moment passés en dehors du cadre professionnel. Je tiens à remercier tout particulièrement trois personnes, Amélie, Débo et Laurie, qui ont été mes plus proches collègues et amies pendant ces années : merci pour vos conseils, votre soutien, votre écoute, et pour tous les délires partagés ensemble ! Sans oublier également tous ceux de passage parmi nous pendant ces années, en particulier Gaëlle et ses folles autruches, Sean, Christine, Thibaut, et Maguelonne.

Je ne voudrais surtout pas oublier Claire Bardel, avec qui j'ai énormément apprécié de travailler et d'échanger, à la fois sur les plans scientifique et personnel, mais également pour m'avoir fait découvrir et intégrée au monde de l'enseignement. Merci également à Vincent Danjean pour son aide sur ALTree. Merci à `./ed` de m'avoir donné l'opportunité d'enseigner au

sein de la formation que j'ai moi-même suivie à l'INSA de Lyon.

Merci à tous mes amis d'avoir été présents : Julie, Manon et Mickaël pour les bonnes soirées et week-ends entre amis, et Julien pour tous ces bons moments musicaux pendant ces dernières années : merci de me faire profiter de ton amitié et de ta créativité ! Je pense aussi à toute la joyeuse bande musicale des OGS et en particulier à Jess, qui transforment le lundi en un des plus beaux jours de la semaine et m'ont permis de m'évader et d'évacuer le stress durant cette dernière année.

Je pense bien sûr également à ma famille : merci à mes parents pour m'avoir toujours soutenue tout au long de mes études dont cette thèse est l'aboutissement. Un grand merci à ma petite soeur Pauline d'avoir consciencieusement relu l'intégralité de ce manuscrit et d'en avoir patiemment relevé les coquilles et autres fautes d'orthographe et de style... Cette relecture m'a été vraiment précieuse !

Enfin, un immense merci à celui qui partage ma vie. Guillaume, merci de m'avoir supportée dans les moments plus difficiles, de m'avoir épaulée pendant ces trois années. Merci de ta présence et de ton soutien indéfectible !

RÉSUMÉ

Au cours de sa vie, une femme sur 9 sera confrontée personnellement à un cancer du sein. Que ce soit pour le cancer du sein sporadique (le plus fréquent) ou pour le cancer du sein familial, il est aujourd'hui indéniable que la génétique influence le risque de développer cette maladie. Bien que de nombreuses mutations causales et polymorphismes de susceptibilité aient déjà été identifiés, une large part de la composante génétique du cancer du sein reste encore à expliquer.

Dans ce contexte, j'ai entrepris d'étudier la variabilité du génome mitochondrial, en essayant de déterminer dans quelle mesure une partie de l'héritabilité manquante du cancer du sein peut être liée à des variants localisés sur le génome mitochondrial. En effet, le génome mitochondrial, présent en plusieurs dizaines voire centaines de copies dans la plupart de nos cellules, est souvent mis de côté lors de l'analyse du génome nucléaire. Ses spécificités, telles que son haploïdie, ont pour conséquence que les études de liaison génétique ne sont pas adaptées à son analyse. De même, jusqu'à très récemment, les puces commerciales disponibles afin de réaliser des études pangénomiques, ou GWAS, couvraient de manière très partielle la variabilité de ce génome. Certains variants de susceptibilité pourraient donc se situer sur le génome mitochondrial sans pour autant avoir été détectés jusqu'ici. De plus, le génome mitochondrial encode pour des gènes majoritairement impliqués dans la production d'ATP de la cellule. Cette production d'ATP génère des composés mutagéniques appelés espèces oxygénées réactives, pouvant contribuer à l'instabilité génomique et à terme à la carcinogénèse. Cette synthèse énergétique est d 'autre part dérégulée dans de nombreux cancers.

Un premier axe de recherche m'a conduit à m'intéresser à une interaction potentielle entre certains variants du génome mitochondrial et du génome nucléaire, en conjonction avec l'exposition à un facteur environnemental lié au style de vie qu'est la consommation d'alcool, et ce dans la population générale. L'absence d'interaction observée dans notre jeu de données, dont la taille est pourtant largement suffisante pour nous conférer la puissance nécessaire, laisse à penser que les résultats antérieurs mettant en évidence de telles interactions ne sont pas robustes.

Je me suis par la suite plus spécifiquement attachée à l'étude de la prédisposition au cancer du sein familial. J'ai tout d'abord étudié dans quelle mesure les variations du génome mitochondrial, et plus spécifiquement les haplogroupes mitochondriaux, peuvent être considérés comme des modificateurs de l'association entre le risque de cancer du sein et certaines mutations causales connues, notamment celles situées sur les gènes *BRCA1* et *BRCA2*. L'haplogroupe T1a1 a été identifié comme modificateur du risque initialement conféré par les mutations pathogènes localisées sur le gène *BRCA2*. Ces résultats sont cohérents avec certaines hypothèses fonction-

nelles basées sur la dérégulation des systèmes de correction des dommages de l'ADN.

Enfin, toujours dans l'objectif de localiser des candidats potentiels permettant d'identifier une partie des variants associés à la proportion non-expliquée de l'héritabilité du cancer du sein, j'ai caractérisé par séquençage à haut débit le génome mitochondrial de plusieurs centaines de femmes diagnostiquées pour un cancer du sein et présentant de forts antécédents familiaux pour cette pathologie, mais n'étant porteuses d'aucune des mutations causales connues à ce jour sur *BRCA1* et *BRCA2*. Plusieurs variants, dont certains non-référencés dans la littérature ni dans les bases de données spécialisées sur la mitochondrie, sont de plus prédits comme dommageables pour le fonctionnement des protéines dont ils affectent la séquence codante. Deux gènes en particulier *MT-ATP6* et *MT-CYB*, sont spécifiquement enrichis à la fois en nombre de variants portés, et de par le nombre d'individus porteurs de ces variants dans notre étude. D'autre part, ces deux gènes codent pour des composants structuraux essentiels de la chaîne respiratoire mitochondriale, la première source de production d'espèces oxygénées réactives de la cellule. L'ensemble des éléments présentés ne constitue pas de preuves formelles de l'implication de ces candidats potentiels dans la susceptibilité du cancer du sein, et des études fonctionnelles sont requises afin de clarifier leur rôle. Il apparaît cependant peu probable qu'une large part de l'héritabilité manquante du cancer du sein puisse être expliquée par l'existence de variants portés par le génome mitochondrial.

L'ensemble du travail réalisé a ainsi contribué à enrichir les connaissances sur les potentielles associations entre les variations du génome mitochondrial et le risque de cancer du sein, en intégrant à la fois des aspects liés aux interactions entre variants génomiques, en tenant compte de certaines expositions environnementales, et en analysant de possibles modifications d'effets liés aux haplogroupes mitochondriaux.

Table des matières

Table des figures

Chapitre 2 **86**

Chapitre 3 **112**

Discussion **144**

Liste des tableaux

INTRODUCTION

I. Notions générales sur le cancer du sein

I.1 Le cancer : une maladie d'origine génomique

Le cancer est une maladie dont les mécanismes déclencheurs trouvent leur origine dans notre génome. Ce dysfonctionnement est causé par une ou plusieurs mutations qui altèrent l'expression de certains gènes impliqués dans la régulation et le contrôle des cellules qui composent notre corps.

Une mutation est une modification de l'information génétique portée par notre génome. Les mutations apparaissent de manière aléatoire le long du génome. Leur fréquence d'apparition peut être augmentée suite à l'exposition à certaines substances dites mutagènes. Parmi les agents mutagènes notoires, on trouve l'exposition aux radiations ionisantes qui peuvent induire entre autres leucémies, lymphomes, mais également des cancers de la thyroïde, du sein et du poumon[1]. On trouve également des produits chimiques tels que les composés de type dioxines, majoritairement synthétisés à partir de procédés de combustion comme l'incinération des déchets ou des feux de forêts, ou encore certains procédés industriels tels que la production de pâte à papier. Cette exposition pourrait être à l'origine d'une augmentation du risque de lymphome hodgkinien et non-hodgkinien, de sarcome, ou de la cancer du foie, du poumon, de la peau, ou de la thyroïde[2]. D'autre part, les mutations intervenant dans les cellules reproductrices ou gamètes seront transmises à la descendance. Les mutations sont ainsi le principal moteur de l'Évolution puisqu'elle favorisent la diversité moléculaire et phénotypique.

Contrairement à l'idée véhiculée par le terme, une mutation n'a au départ aucune connotation positive ou négative, puisque c'est un processus essentiellement aléatoire. Elles n'ont cependant pas toutes les mêmes conséquences. Certaines apparaissent dans des gènes ayant une fonction critique dans la mise en oeuvre et le contrôle des processus biologiques de notre corps. Parmi ces gènes dont le rôle est primordial, on trouve trois catégories qui sont les proto-oncogènes, les gènes suppresseurs de tumeurs, et les gènes de réparation de l'ADN.

Un gène est appelé proto-oncogène s'il a le potentiel de stimuler le développement d'une tumeur. Dans une cellule normale, les fonctions des proto-oncogènes sont relatives, entres autres, à la division et à la prolifération cellulaire, ainsi qu'à la différenciation[3]. Un proto-oncogène peut ainsi être activé suite à l'apparition d'une mutation qui modifie son expression ou sa régulation ; à la suite de son activation, un proto-oncogène est alors appelé oncogène. Les oncogènes vont alors favoriser localement le développement tumoral en stimulant, entre autres, la

synthèse de facteurs de croissance, de facteurs de transcription, et la prolifération des cellules tumorales. Les proto-oncogènes peuvent être exprimés dans une large variété de tissus et types cellulaires. C'est pourquoi on les retrouve mutés dans de nombreux types de cancers. Ainsi, les trois gènes de la famille RAS, appelés respectivement *K-RAS*, *H-RAS*, et *N-RAS*, semblent être des acteurs prépondérants au cours du développement tumoral. *K-RAS* est muté de manière somatique dans approximativement 90% des cas de cancer du pancréas, mais aussi dans 50% des cancers du colon, et dans 30% des adénocarcinomes du poumon[4]. *H-RAS* a été observé muté dans des tumeurs de la thyroïde, de la peau, des cervicales, ou encore de la tête et du cou[4].

Les gènes suppresseurs de tumeurs ont quant à eux le rôle de gardien de l'intégrité des processus cellulaires. Ces gènes sont normalement exprimés en cas de stress cellulaire provoqué par exemple par des dommages sur l'ADN. Ils interviennent sur le contrôle de la croissance cellulaire. Dans certains cas, par exemple si les dégâts occasionnés à une cellule sont trop importants, certains ont la capacité d'induire le processus de mort cellulaire programmée, appelé apoptose. Le gène suppresseur de tumeur le plus connu aujourd'hui est sûrement p53. Dans un contexte normal, p53 est exprimé en réponse à un stress cellulaire. Il stoppe alors la croissance cellulaire, le temps que d'autres mécanismes (de réparation de l'ADN par exemple) se mettent en place. p53 peut également induire l'apoptose. Les mutations de p53 peuvent altérer ses fonctions. Lorsque p53 ne peut plus exercer son rôle, les dégâts sur l'ADN cellulaire s'accumulent et conduisent à un phénotype cancéreux[5]. p53 a été observé muté dans environ 10% des cancers hématopoïétiques, et dans 50% à 70% des cancers de l'ovaire, de la tête et du cou, et colorectal[6].

L'ADN de nos cellules est soumis à des attaques permanentes qui lui occasionnent des dommages. Des mécanismes de réparation de l'ADN existent et peuvent être classés en quatre catégories[7] :

- Réparation par excision d'une base, ou *Base Excision Repair* : système de réparation qui prend en charge les dommages les plus simples et les moins étendus.
- Réparation par excision d'un nucléotide, ou *Nucléotide Excision Repair* : système de réparation qui prend en charge les dommages affectant la structure de la double-hélice d'ADN.
- Réparation des mésappariements, ou *Mismatch Repair* : système qui corrige les erreurs de réplication.
- Réparation des cassures double-brins, ou *Double Strand Break Repair* : système de réparation intervenant lorsque les cassures occasionnées affectent les deux brins d'ADN. Ce type de réparation peut intervenir via différents processus tels que celui de la recombinaison homologue, ou celui de la ligation non-homologue.

Ces mécanismes sont essentiels au maintien de l'intégrité de l'ADN cellulaire, et font intervenir de très nombreux gènes. Leur altération conduit à une accumulation des erreurs sur l'ADN, pouvant conduire à une inactivation d'autres gènes, ou à une instabilité génomique à l'origine du développement tumoral.

L'altération des gènes appartenant aux trois catégories décrites ci-dessus peut conduire au développement d'un cancer. Un cancer peut se manifester physiquement de plusieurs manières.

Les cancers caractérisés par une « tumeur solide » peuvent se développer dans la plupart des tissus, et représentent 90% des cancers humains. On peut les classer en deux types : les *carcinomes*, qui se forment à partir de cellules dites épithéliales (peau, glandes, muqueuses), et les *sarcomes* qui se forment à partir de cellules des tissus conjonctifs (tissus dits de « soutien », comme les os et les cartilages). Les cancers sanguins se développent à partir de cellules situées dans un milieu liquide. Les cancers du sang et de la moelle osseuse sont appelés leucémies, alors que les lymphomes sont des cancers qui affectent les ganglions et le système lymphatique. Enfin, les cancers métastatiques apparaissent à la suite d'un cancer primaire, alors que des cellules cancéreuses ont pénétré le système sanguin ou lymphatique et ont colonisé d'autres organes distants, tels que les os, le foie, les poumons, ou le cerveau.

I.2 Statistiques épidémiologiques sur le cancer du sein

Le cancer du sein est une maladie essentiellement féminine, mais elle peut se développer également chez les hommes. Cependant, la proportion des cancers du sein diagnostiqués chez des hommes reste très faible, de l'ordre de 0.5% à 1%[8]. L'incidence moyenne du cancer du sein masculin est de 1/100 000 en Amérique du nord et en Europe de l'ouest[9]. Ce taux d'incidence aurait tendance à augmenter[10]. Le travail de thèse qui a été réalisé et qui est présenté dans ce manuscrit est axé sur le cancer du sein en tant que pathologie féminine.

Le Centre International de Recherche contre le Cancer surveille de manière continue les tendances concernant l'incidence, la mortalité et le dépistage de tous les types de cancers à l'échelle mondiale. Les résultats de ce projet, appelé GLOBOCAN, ont été publiés en 2008 et en 2012[11]. Ce sont les données les plus récentes et les plus complètes dont on dispose pour étudier tous types de cancers dans le monde.

D'après le Centre International de Recherche contre le Cancer, on estime que 1 676 633 nouveaux cas de cancers du sein ont été diagnostiqués dans le monde en 2012 , ce qui représente 11,9% de tous les cancers diagnostiqués, tous sexes confondus. En 2012, 6.3 Millions de femmes vivent alors qu'un cancer du sein leur a été diagnostiqué dans les cinq dernières années. Entre 2008 et 2012, l'incidence du cancer du sein à l'échelle mondiale a augmenté de 20%, alors que la mortalité a augmenté de 14%, tous pays confondus. Le cancer du sein est aussi la première cause de mortalité chez les femmes, avec 522 000 décès en 2012.

En 2012, les taux d'incidence mondiaux les plus élevés sont observés en Amérique du nord, en Europe occidentale, en Australie. Les taux de mortalité les plus élevés sont quant à eux observés en Afrique subsaharienne, au Moyen-Orient. Cependant, le taux d'incidence des pays en voie de développement va probablement rapidement augmenter les prochaines décennies. D'autre part, le cancer du sein est une des premières causes de mortalité due au cancer dans ces pays, les avancées cliniques les plus récentes et les traitements associés ne s'y démocratisant pas encore[11].

Historiquement, les femmes d'origine asiatique et des îles pacifiques ont le taux d'incidence pour le cancer du sein le plus faible, puis viennent les hispaniques, les Afro-américaines. Enfin, les femmes caucasiennes présentent le taux d'incidence le plus élevé[12]. Cependant cette répartition évolue, et des distinctions semblent apparaître entre les différentes éthnies asiatiques.

Le taux d'incidence chez les Japonais et les Philippins semblent augmenter bien plus vite que celui des Chinois et des Coréens[13]. Les taux d'incidences des Caucasiens et des Afro-américains montrent une tendance nette à se rejoindre depuis les dernière années[12].

Le cas des immigrants asiatiques venus s'installer aux États-Unis est frappant : bien qu'ayant un taux d'incidence très bas dans leur pays natal, celui-ci augmente très rapidement en seulement quelques générations jusqu'à devenir semblable à celui des Caucasiens. De plus, les femmes ayant émigré aux États-Unis et y vivant depuis au moins 10 ans ont un risque de cancer du sein plus élevé de 80% en comparaison des femmes arrivées plus récemment[14]. Cela démontre l'importance du style de vie et le rôle crucial de l'environnement dans le risque de cancer du sein.

En France, le taux d'incidence du cancer du sein à tendance à augmenter, en partie à cause des campagnes de dépistage systématique mises en place. Sa valeur en 2012 était de 88 pour 100 000 femmes par an, année durant laquelle une très légère baisse du taux d'incidence a été enregistrée. A contrario, le taux de mortalité est en baisse, avec une valeur en 2012 de 15.7 pour 100 000 femmes par an. L'âge moyen du diagnostic est de 63 ans. Pour un cancer diagnostiqué entre 1989 et 2004, le taux de survie est de 97% à 1 an, 86% à 5 ans, et 76% à 10 ans (Chiffres de l'Institut National du Cancer, juillet 2013).

I.3 Classifications du cancer du sein

I.3 .1 Classifications clinique et moléculaire

Le cancer du sein se manifeste sous la forme d'une tumeur solide. Plus de 95% des cancers du sein sont des adénocarcinomes, soit des tumeurs qui se développent à partir des cellules de type épithélial (*-carcinome*) qui constituent les tissus des glandes (*adeno-*) mammaires. Une tumeur mammaire se caractérise au moment du diagnostic et au cours de son évolution par ses dimensions et sa localisation. Les cancers du sein peuvent se développer dans les canaux de la glande mammaire - cancer canalaire - ou bien dans les lobules - cancer lobulaire - (voir Fig. 1). Lorsque la tumeur est située dans ces deux régions, on parle de carcinome *in situ*. À l'inverse, lorsque les tissus périphériques sont atteints, le cancer est qualifié d'infiltrant. On parle d'envahissement ganglionnaire lorsque les ganglions axillaires, situés sous le bras, sont également touchés par la tumeur. Enfin, le cancer du sein est dit métastatique lorsque des cellules cancéreuses se sont disséminées depuis la tumeur mammaire jusque dans d'autres organes.

Les tumeurs du sein présentent une forte variabilité au niveau moléculaire. En effet, en plus des critères morphologiques tels que la taille et la localisation, une tumeur peut être décrite du point de vue moléculaire. De manière concrète, cela revient à établir si certains gènes sont exprimés et si certaines protéines sont présentes au niveau de la tumeur. Il est important de savoir si ces gènes et protéines sont exprimés ou présents dans la tumeur dans la mesure où leur présence ou leur absence permet de classer les tumeurs par sous-types, et de déterminer quels traitements ont le plus de chance d'être efficaces.

L'étude des profils d'expression de gènes dans la tumeur ont permis de constituer une classification moléculaire des cancers du sein par sous-type. Un des critères principaux de cette classification est l'expression ou non des gènes codant pour les récepteurs hormonaux : si la

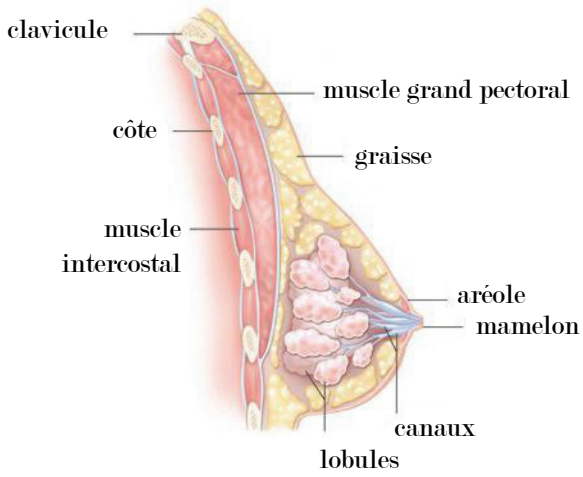


Figure 1 – Schéma d'un sein en coupe

tumeur exprime les gènes *ER* et/ou *PR* (on dit que alors que la tumeur est ER+ et/ou PR+), alors les cellules tumorales présenteront à leur surface des récepteurs aux oestrogènes et à la progestérone. Un traitement par hormonothérapie pourra alors être envisagé. Un autre critère primordial dans cette classification est la recherche de l'expression *Human Epidermal growth factor Receptor 2* , ou HER2. Une tumeur HER+ possèdera à la surface de ses cellules des récepteurs à cette hormone de croissance. Ces biomarqueurs ont de plus une signification prédictive - comment évoluera la tumeur ? - et pronostique - le(s) traitement(s) administrés seront-ils efficaces ? La classification actuelle, bien qu'imparfaite, est constituée par 5 sous-types principaux : Luminal A, Luminal B, HER2+, *Basal-Like*, et Triple-Négatif.

Les sous-types Luminal A et Luminal B expriment des gènes codant pour des protéines caractéristiques des cellules épithéliales situées dans la lumière des canaux ou dans les lobules des glandes mammaires. Les tumeurs de type Luminal A sont généralement ER+ et/ou PR+, et HER2-. Les tumeurs de type Luminal B sont généralement RE+ et/ou PR+, et HER2+. Les tumeurs Luminales A coexpriment plus fréquemment ER et PR que les tumeurs Luminales B. Elles sont également associées à un meilleur pronostic[15]. Les tumeurs Luminales B présentent plus souvent une mutation du gène suppresseur de tumeur p53. L'hormonothérapie appliquée à des tumeurs Luminal B est souvent peu efficace et requiert plutôt une chimiothérapie.

Le sous-type HER2+ se caractérise par une surexpression du gène *HER2*. Cet oncogène appartient à la famille des gènes codant pour les récepteurs aux facteurs de croissance. Il est surexprimé dans environ 20% des cancers de sein[16,17]. La surexpression de *HER2* et de la protéine correspondante est associée à une augmentation de la prolifération et de la motilité cellulaire, une augmentation de l'angiogénèse et du potentiel invasif de la tumeur, et une diminution de l'apoptose[18].

Les sous-types *Basal-Like* (BL) et Triple Négatif (TN) se ressemblent. Le sous-type Triple

Négatif (TN) est constitué par des tumeurs ER-, PR- et HER2-, il représente environ 15% des cancers du sein[19]. Les tumeurs de sous-type sont souvent de grade plus élevé, et associées à un phénotype plus agressif, ainsi qu'à un plus mauvais pronostic[20,21]. Aujourd'hui il n'existe pas de thérapie ciblée dédiée au traitement des tumeurs TN. Tout comme celles du sous-type TN, les tumeurs BL n'expriment que peu ou pas de récepteurs hormonaux (ER-, PR-, HER2-). Cependant, les profils d'expression génique de ces deux sous-types sont différents. Les tumeurs BL sont caractérisées par un profil d'expression dans lequel on retrouve des gènes exprimés habituellement au sein de cellules basales ou myoépithéliales de tissus normaux du sein[22].

Aujourd'hui cette classification des cancers du sein reste imparfaite, et évoluera certainement. Elle reste perpétuellement débattue, et bien que les sous-types décrits ci-dessus soient peu remis en question, d'autres sous-types apparaissent, tels que le sous-type *Claudin-Low* qui serait associé à un phénotype de récurrence et à l'apparition de métastases[23]. Cette classification des tumeurs en sous-types permet d'appliquer dès le diagnostic le principe de médecine personnalisée, en administrant les traitements les plus susceptibles de fonctionner.

I.3 .2 Traitements classiques et thérapies ciblées

Une tumeur étant formée par un ensemble de cellules sur lesquelles l'organisme a perdu le contrôle, il existe peu de moyens d'agir sur elle. Les principaux traitements actuels sont la chirurgie, la radiothérapie, la chimiothérapie, l'hormonothérapie, ainsi que les thérapies ciblées.

La chirurgie La chirurgie a pour but de faire l'ablation de la tumeur et d'un maximum de cellules tumorales. Selon le stade de développement de la tumeur et son niveau d'infiltration, deux types de chirurgie peuvent être effectués. Pour une chirurgie dite conservatrice, c'est uniquement la tumeur et éventuellement quelques tissus périphériques qui sont enlevés, on parle de *tumorectomie*. Ce type de chirurgie peut être précédé par d'autres traitements dans le but de diminuer la taille de la tumeur afin de rendre son extraction plus facile. Lorsque la tumeur s'est infiltrée plus profondément dans les tissus, il est nécessaire d'effectuer l'ablation du sein entier, on parle alors de *mastectomie*. L'ablation de certains ganglions à proximité de la tumeur peut également être effectuée.

La radiothérapie La radiothérapie est un traitement visant à détruire localement les cellules tumorales par irradiation. La radiothérapie externe est appliquée depuis l'extérieur du corps, alors que la curiethérapie, moins répandue, consiste à irradier les cellules de l'intérieur en plaçant un implant émettant des radiations à proximité de la tumeur. Dans les deux cas, les radiations appliquées sont dites ionisantes, l'énergie qu'elles portent est transformée en dommages thermiques et chimiques sur la cellule. L'ADN de la cellule subit alors une très forte quantité de dommages. La cellule peut alors rentrer en apoptose si les mécanismes associés sont encore actifs dans la tumeur. Dans le cas contraire, les dégâts occasionnés à l'ADN et aux membranes cellulaires vont conduire à la mort cellulaire lorsque la cellule tentera de se diviser.

La chimiothérapie La chimiothérapie vise à ralentir le développement de la tumeur en empêchant les cellules de se diviser et dans une moindre mesure en induisant leur mort. Les substances chimiques administrées ont pour rôle de bloquer la division cellulaire en agissant par divers mécanismes directs ou indirects sur l'ADN cellulaire, sur sa synthèse, sa disponibilité, ou sa réplication. La chimiothérapie est efficace sur une tumeur lorsque celle-ci est dans une phase de division cellulaire rapide. Plus les cellules se divisent, plus la chimiothérapie sera efficace. Cela constitue également son plus gros inconvénient. En effet, la chimiothérapie étant administrée à l'échelle de l'organisme entier, toutes les cellules à développement rapide subissent également les conséquences du traitement, et notamment les cellules capillaires, les cellules épithéliales intestinales, et les cellules sanguines. C'est pourquoi les principaux effets secondaires du traitement sont la perte de cheveux, des difficultés d'alimentation, la baisse de l'immunité et l'anémie.

L'hormonothérapie L'hormonothérapie est un traitement visant à bloquer la stimulation de la croissance de la tumeur. En effet, certaines tumeurs sont sensibles aux hormones féminines qui stimulent leur croissance. Chez les femmes atteintes d'un cancer du sein, le traitement administré a pour but de diminuer la production d'oestrogènes ou de limiter la sensibilité de la tumeur à ces hormones. De manière concrète, cela se traduit par l'administration d'hormones de substitution qui viennent se fixer sur les récepteurs tumoraux aux oestrogènes, ce qui bloque la stimulation. On peut également administrer des composés analogues aux hormones hypothalamiques qui exercent un rétrocontrôle négatif sur la production d'oestrogènes et limitent leur synthèse.

Les thérapies ciblées Au delà de la caractérisation globale de la tumeur du point de vue moléculaire, les techniques de génomique actuelles permettent de séquencer tous ou une partie des gènes de la tumeur afin de détecter certaines anomalies génomiques précises ayant un rôle causal dans le développement ou dans la progression du cancer du sein. En étudiant les dysfonctionnements cellulaires conduisant à l'apparition du cancer, on a ainsi pu caractériser ces anomalies. On peut aujourd'hui compenser les effets de certaines de ces anomalies par des traitements spécifiques, c'est ce qu'on appelle les thérapies ciblées. Les thérapies ciblées sont l'exemple même de la médecine personnalisée. Par exemple, les anticorps monoclonaux connus sous le nom de *trastuzumab* ont constitué la première thérapie ciblant les tumeurs HER2+. D'autres inhibiteurs sont actuellement en développement afin de cibler une voie de signalisation dérégulée dans les cancers de type ER+ et HER2+, la voie PI3K-AKT-mTOR[24]. De plus, les différents sous-types tumoraux présenteraient des altérations spécifiques de cette voir métabolique[25]. Des thérapies ciblées sont également en cours de développement afin de traiter les tumeurs présentant une amplification des gènes de la famille *FGFR*. Pour le sous-type Luminal B, généralement associé à un mauvais pronostic, la fréquence d'amplification de *FGFR1* atteint 27%[26]. La thérapie ciblée la plus avancée du point de vue du développement clinique est celle des inhibiteurs de la tyrosine kinase.

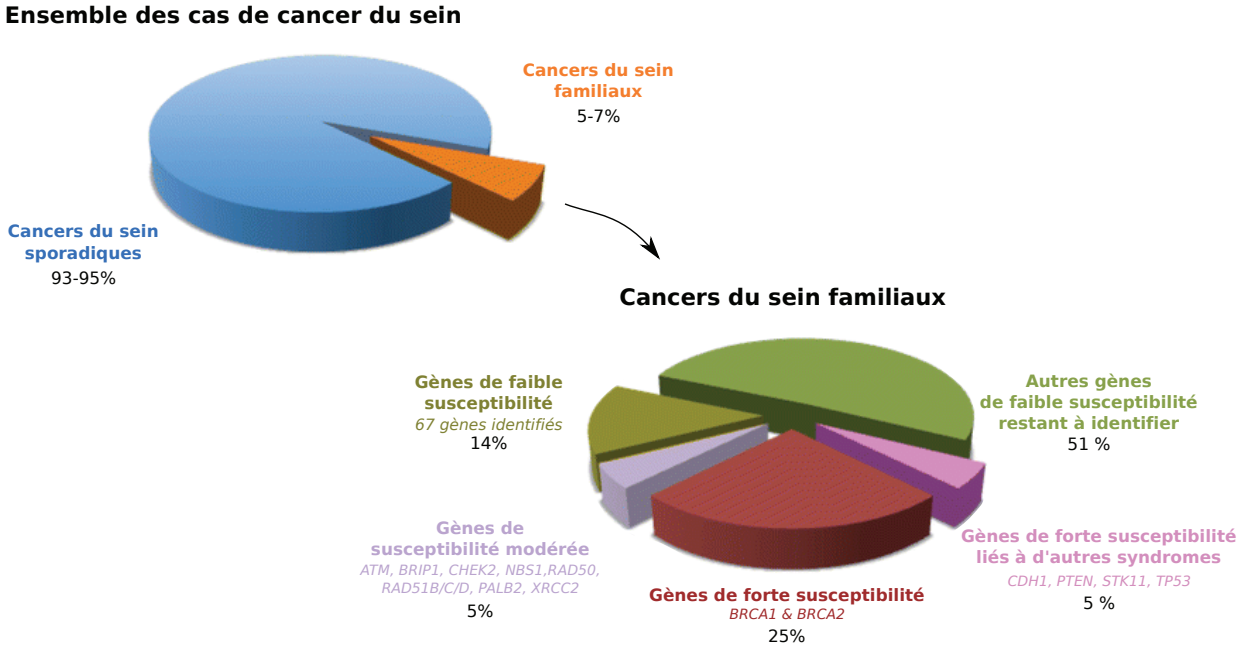
I.4 Facteurs de risque pour le cancer du sein

De très nombreux facteurs ont été identifiés comme pouvant modifier le risque de cancer du sein chez les femmes. Les facteurs de risque se classent en deux catégories : les facteurs dits non génétiques, qui regroupent le style de vie, l'exposition environnementale, l'alimentation ; et les facteurs génétiques, observés dans le génome de l'individu.

I.4 .1 Composante génétique du cancer du sein

On distingue deux catégories de cancers du sein, les cancers du sein sporadiques, et les cancers du sein familiaux. Alors que les cancers du sein sporadiques sont des cas « isolés », on parle de cancer du sein familial lorsque plusieurs cas de cancers du sein sur plusieurs générations sont observés au sein d'une même famille. L'âge au diagnostic des cas de cancer du sein familiaux est général plus faible, avant 50 ans. Dans ces familles, cette agrégation de cas parmi un ensemble d'individus apparentés laisse à penser qu'un facteur génétique intervient dans la susceptibilité. Les cancers du sein familiaux ne représentent qu'un faible pourcentage de l'ensemble des cas de cancers du sein, entre 5% et 7% (Figure 2). En se basant sur les chiffres de l'incidence mondiale du cancer du sein pour l'année 2012[11], cela représenterait cependant environ 100 000 nouveaux cas par an.

Figure 2 – Composante génétique du cancer du sein - *Adaptation de Melchor et al. Human Genetics, 2013*[27]

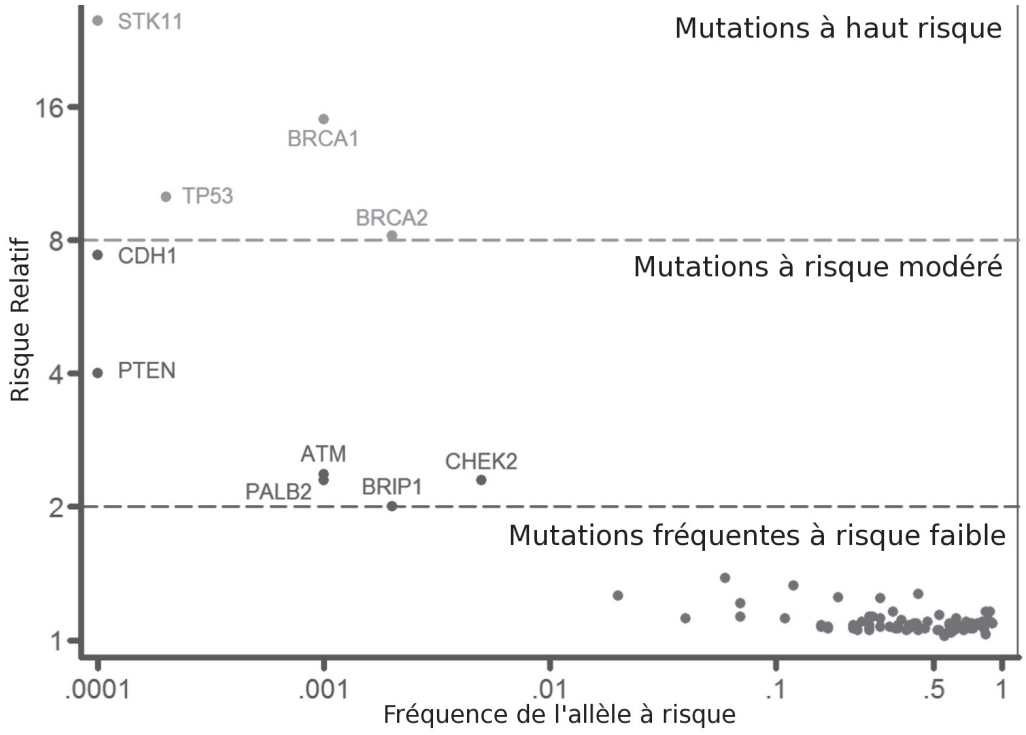


L'observation de ces familles pour lesquelles le risque de cancer du sein est visiblement plus élevé que dans la population générale a conduit à rechercher en plus des causes environ-

nementales, une cause génétique. Aujourd'hui, on estime qu'environ 50 % de la composante génétique du cancer est expliquée, c'est à dire qu'on a identifié environ la moitié des facteurs de prédisposition génétique (Figure 2).

Les facteurs de prédisposition génétiques au cancer du sein sont des mutations ou des polymorphismes. Ces variants génomiques sont plus ou moins fréquents dans la population, et on un effet plus moins fort. Généralement, les variants dont l'effet est le plus important sont les plus rares. On les classe généralement en trois classes : les variants à effet faible, les variants dont l'effet est modéré, et les variants de forte susceptibilité. La figure 3 synthétise l'ensemble des variants et des gènes identifiés pour affecter la susceptibilité au cancer du sein en fonction de leur fréquence allélique et de leur effet estimé. Cette représentation illustre parfaitement à quel point la composante génétique du risque de cancer du sein est complexe, et à quel point cette maladie est multifactorielle.

Figure 3 – Ensemble des loci de susceptibilité au cancer du sein identifiés représentés en fonction de leur fréquence allélique et de la force de leur effet sur le risque de cancer du sein



Adaptation de Ghoussainni et al. The American Journal of Pathology, 2013[28]

Une proportion importante des familles présentant une histoire familiale de cancer du sein allant de 15% à 40 %[29], possèdent une mutation pathogène dans *BRCA1* ou *BRCA2* qui altère la structure ou la fonction des protéines encodées par ces gènes. Ces deux gènes sont des acteurs majeurs de la réparation de l'ADN. Ainsi, une femme porteuse d'une mutation pathogène sur *BRCA1* a 65% à 85% de chances d'avoir un cancer du sein au cours de sa vie[30,31], et cette mutation augmente le risque de cancer de l'ovaire également. De même, une femme porteuse

d'une mutation pathogène sur *BRCA2* a un risque de développer un cancer du sein au cours de sa vie d'environ 40 % à 85%. Aujourd'hui, plus de 100 mutations pathogènes ont été identifiées sur *BRCA1*, et chaque famille ayant des antécédents de cancers du sein présente généralement une mutation unique. Cependant, certaines mutations sur *BRCA1/2* ont été observées associées à un effet fondateur, c'est à dire qu'elles étaient portées par quelques individus qui, pour des raisons géographiques, sociales, religieuses ou autres, se sont retrouvés à l'origine de la formation d'une nouvelle communauté sans qu'il y ait eu de brassage génétique externe. Les polymorphismes portés par les individus fondateurs ont donc une fréquence plus élevée dans cette communauté que dans la population générale. C'est le cas par exemple, de la communauté des juifs ashkénazes. Cette communauté regroupant les individus de confession juive d'Europe a des origines génétiques diverses provenant entre autres du Moyen-Orient, du Caucase et de Mésopotamie. Leur migration en Europe de l'est et en Europe Occidentale remonte au Ier siècle de notre ère, et a entraîné un effet fondateur pour certaines mutations. En effet, les 3 mutations suivantes sont observées relativement fréquemment dans cette communauté : une insertion d'une base C en position 5382 sur *BRCA1* (notée 5382insC), et une délétion de deux bases AG en position 185 de *BRCAC1* également (notée 185delAG), ainsi qu'une délétion d'une base T en position 6147 sur *BRCA2* (notée 6147delT). Alors que la probabilité d'être porteur de l'une de ces trois mutations est d'environ 0.02 % dans la population générale, elle a été estimée à 2.5% chez les juifs ashkénazes[32–34]. De nombreuses autres mutations spécifiques de certaines origines géographiques ou communautés ont ainsi été identifiées[35]. La délétion de 11 paires de bases en position 3600 de *BRCA1* (notée 3600del11), semble être spécifique de certaines familles originaires du nord de la France.

Les variants à effet modéré sont principalement des variants localisés dans des gènes impliqués également dans la réparation de l'ADN. Les protéines encodées par ces gènes sont pour la plupart des partenaires de BRCA1 et BRCA2. Ainsi, des variants dans les gènes *ATM*, *BRIP1*, *CHEK2*, *NSB1*, *RAD50*, *RAD51B*, *RAD51C*, *RAD51D*, *PALB2* et *XRCC2* multiplient le risque de cancer du sein par un facteur compris entre 2 et 3.

Le gène suppresseur de tumeur *TP53*, encodant le facteur de transcription p53, a été identifié comme potentiel gène candidat dans le cadre de l'étude des causes génétiques du syndrome de Li-Fraumeni[36,37], un syndrome rare se manifestant par une incidence élevée de cancer du sein uni- et bilatéral, de sarcomes des tissus mous chez l'enfant, et d'ostéosarcome chez les jeunes adultes. La moitié des cancers dus à ce syndrome se manifestent avant l'âge de 30 ans. Plusieurs mutations constitutionnelles ont été décrites[38,39], principalement entre les exons 5 et 8 du gène, région d'interaction de p53 avec l'ADN, toutes ont pour conséquence d'inactiver le rôle protecteur de p53 qui est de stopper le cycle cellulaire en cas de dommages occasionnés sur l'ADN, afin d'éviter que la cellule ne se multiplie. Aujourd'hui, il semblerait que l'inactivation de p53 favorise également l'invasion, le développement métastatique, la prolifération et la survie cellulaire[40].

À la suite de l'identification de *BRCA1/2* comme gènes de susceptibilité au cancer du sein, la communauté scientifique a cherché à caractériser ces protéines, leur rôle, et les protéines partenaires avec lesquelles elles interagissent *in vivo*. Cela a mené à l'identification de PALB2 - pour *PArtner and Localizer of BRCA2* - qui encode une protéine dont la présence est requise

pour que BRCA2 puisse assurer son rôle dans la recombinaison homologue et la réparation des cassures double-brins. L'analyse des polymorphismes sur *PALB2*[41] a conduit à l'identification de SNPs, majoritairement non-sens, qui augmentent de manière significative le risque de cancer du sein - OR = 2.3 (IC95% 1.4–3.9, p-value = 0.0025).

Les protéines encodées par gènes *CHEK2* et *ATM* sont deux partenaires de p53. CHECK2 est également un partenaire direct de BRCA1. En tant que tel, les mutations constitutionnelles sur ce gène ont été étudiées par une approche gène candidat. La mutation 1100delC qui inactive CHECK2, a été observée plus fréquemment chez des individus atteints d'un cancer du sein familial, mais ne portant pas de mutation sur *BRCA1/2*[42].

ATM est un gène codant pour un protéine kinase également impliquée dans le contrôle et la réparation des cassures double-brins. Ce gène a été initialement identifié dans le cadre de la recherche de la cause génétique de l'ataxie télangiectasie, un syndrome sévère se caractérisant par des atteintes neurologiques et immunitaires. Ce syndrome est une maladie génétique autosomique récessive, et est caractérisé par des mutations homozygotes localisées de manière homogène sur le gène *ATM*. Plus de 500 mutations ont été répertoriées. Il a été observé au sein des familles des individus atteints que les femmes porteuses hétérozygotes d'une de ces mutations avaient un risque plus élevé de cancer du sein que le reste de la population générale[43]. Dans l'étude initiale ayant mis en évidence cette observation, les 12 mutations identifiées à l'origine de l'ataxie télangiectasie étaient homozygotes. Cependant, sans pour autant que cela induise les symptômes de l'ataxie télangiectasie, porter une de ces mutations de manière hétérozygote multiplie par 2.37 (IC95% 1.51–3.78, p-value = 0.0003) le risque de développer un cancer du sein.

BRIP1 a tout d'abord été identifié dans le cadre de l'étude de l'anémie de Fanconi, une maladie génétique rare se caractérisant par des atteintes neurologiques, des anomalies de croissance, des malformations congénitales, et de sévères atteintes hématologiques allant jusqu'à la leucémie. Ce gène, codant pour une ADN hélicase - enzyme capable d'ouvrir la double hélice d'ADN afin que d'autres protéines, telles que celles impliquées dans la réparation de l'ADN, puissent effectuer leur rôle - qui est un également un partenaire de *BRCA1* dans la réparation de l'ADN. *BRIP1* a donc fait l'objet d'une étude gène candidat, et plusieurs mutations constitutionnelles ont été associées avec une augmentation du risque de cancer du sein[44,45]. Porter une de ces mutations multiplierait ce risque par 2.0 (IC95% 1.2-3.2, p=0.012). La majorité des mutations ainsi identifiées augmentent le risque de cancer du sein lorsqu'elles sont hétérozygotes, et causent l'anémie de Fanconi lorsqu'elles sont homozygotes.

Certaines mutations sur le gène suppresseur de tumeurs *STK11* provoquent le syndrome de Peutz-Jeghers, atteinte autosomale dominante, qui se manifeste par des atteintes gastro-intestinales ainsi que par un risque accru de cancers, notamment de cancer du sein[46]. Les cas diagnostiqués pour ce syndrome ont un risque de cancer du sein multiplié par un facteur d'au moins 15[47].

Le gène *CDH1* code pour une protéine appartenant à la famille des cadherines, des protéines impliquées dans l'adhésion des cellules par jonctions dépendantes au calcium. Les mutations pathogènes sur ce gène provoquent en premier lieu un risque très élevé de cancer gastrique

diffus, mais prédisposent également au cancer du sein lobulaire[48,49], dont le risque est multiplié par un facteur 8 environ.

De nombreux polymorphismes d'une seule paire de base, pouvant être fréquent dans la population générale ont été identifiés dans notre génome nucléaire[50]. Ces 72 loci sont présentés dans la Table 1. 17 d'entre eux augmentent également le risque de développer d'autres cancers (cancers de l'ovaire et de la prostate notamment). Alors que les altérations à forte pénétrance affectent généralement les séquences nucléotidiques dans leur région codante, et ce en particulier dans les gènes impliqués dans la voie de réparation de l'ADN, ces polymorphismes sont majoritairement synonymes, et sont situés dans des régions non codantes. Ils agiraient plus vraisemblablement au niveau de la régulation des gènes au sein de multiples voies de signalisation.

Table 1 – Polymorphismes affectant la susceptibilité au cancer du sein identifiés par des études pangénomiques

Région	SNP	Gène probablement lié	Année	OR par allèle	p-value	Fréquence de l'allèle à risque
1p11	rs11249433	*NOTCH2/FCGR1B*	2009	1.16 (1.09-1.24)	7×10^{-10}	0.39
1p13	rs11552449	*TPN22/BCL2L15*	2013	1.07 (1.04-1.09)	1.8×10^{-8}	0.17
1p36	rs616488	*PEX14*	2013	0.94 (0.92-0.96)	2.0×10^{-10}	0.33
1q32	rs4245739	*MDM4*	2013	1.14 (1.10-1.18)	2.1×10^{-12}	0.26
1q32	rs6678914	*LGR6*	2013	1.10 (1.06-1.13)	1.4×10^{-8}	0.59
2p24	rs12710696	-	2013	1.10 (1.06-1.13)	1.4×10^{-8}	0.36
2q14	rs4849887	-	2013	0.91 (0.88-0.94)	3.7×10^{-11}	0.10
2q31	rs2016394	*METAP1D*	2013	0.95 (0.93-0.97)	1.2×10^{-8}	0.48
2q31	rs1550623	*CDCA7*	2013	0.94 (0.92-0.97)	3.0×10^{-8}	0.16
2q33	rs1045485	*CASP8*	2007	0.88 (0.84-0.92)	1.1×10^{-7}	0.13
2q33	rs10931936	*CASP8*	2007	0.88 (0.82-0.94)	1.1×10^{-7}	0.26
2q35	rs13387042	*IGFBP2/IGFBP5/TPN2*	2007	1.20 (1.14-1.26)	1×10^{-13}	0.49
2q35	rs16857609	*DIRC3*	2013	1.08 (1.06-1.10)	1.1×10^{-15}	0.26
3p24	rs4973768	*SLC4A7/NEK10*	2009	1.11 (1.08-1.13)	4.1×10^{-23}	0.46
3p24	rs12493607	*TGFBR2*	2013	1.06 (1.03-1.08)	2.3×10^{-8}	0.35
3p26	rs6762644	*ITPR1/EGOT*	2013	1.07 (1.04-1.09)	2.2×10^{-12}	0.40
4q24	rs9790517	*TET2*	2013	1.05 (1.03-1.08)	4.2×10^{-8}	0.23
4q34	rs6828523	*ADAM29*	2013	0.90 (0.87-0.92)	3.5×10^{-16}	0.13
5p12	rs10941679	*MRPS30/HCN1*	2008	1.19 (1.13-1.26)	1×10^{-11}	0.25
5p12	rs9790879	*MRPS30/HCN1*	2008	1.10 (1.03-1.17)	1×10^{-11}	0.40
5p15	rs10069690	*TERT/CLPTM1L*	2011	1.18 (1.13-1.25)	1.0×10^{-10}	0.30
5q11	rs889312	*MAP3K1/MEIR3*	2007	1.13 (1.10-1.16)	1×10^{-15}	0.28
5q11	rs10472076	*RAB3C*	2013	1.05 (1.03-1.07)	2.9×10^{-8}	0.38
5q11	rs1353747	*PDE4D*	2013	0.92 (0.89-0.95)	2.5×10^{-8}	0.10
5q33	rs1432679	*EBF1*	2013	1.07 (1.05-1.09)	2.0×10^{-14}	0.43
6p23	rs204247	*RANBP9*	2013	1.05 (1.03-1.07)	8.3×10^{-9}	0.43
6p25	rs11242675	*FOXQ1*	2013	0.94 (0.92-0.96)	7.1×10^{-9}	0.39
6q14	rs17530068	-	2012	1.12 (1.08-1.16)	1.1×10^{-9}	0.22
6q25	rs3757318	*ESR1*	2009	1.21 (1.13-1.31)	2×10^{-15}	0.07
6q25	rs2046210	*ESR1*	2009	1.11 (1.07-1.16)	3.7×10^{-9}	0.34
7q35	rs720475	*ARHGEF5/NOBOX*	2013	0.94 (0.92-0.96)	7.0×10^{-11}	0.25
8p12	rs9693444	-	2013	1.07 (1.05-1.09)	9.2×10^{-14}	0.32
8q21	rs6472903	-	2013	0.91 (0.89-0.93)	1.7×10^{-17}	0.18
8q21	rs2943559	*HNF4G*	2013	1.13 (1.09-1.17)	5.7×10^{-15}	0.07
8q24	rs13281615	*MYC*	2007	1.08 (1.05-1.11)	5×10^{-12}	0.40
8q24	rs1562430	*MYC*	2010	1.17 (1.10-1.25)	5×10^{-12}	0.40
8q24	rs11780156	*MIR1208*	2013	1.07 (1.04-1.10)	3.4×10^{-11}	0.16
9p21	rs1011970	*CDKN2A/B*	2010	1.09 (1.04-1.14)	2.5×10^{-8}	0.17

Région	SNP	Gène probablement lié	Année	OR par allèle	p-value	Fréquence de l'allèle à risque
9q31	rs865686	*KLF4/RAD23B*	2011	0.89 (0.85- 0.92)	1.7×10^{-10}	0.39
9q31	rs10759243	-	2013	1.06 (1.03-1.08)	1.2×10^{-8}	0.39
10p12	rs7072776	*MLLT10/DNAJC1*	2013	1.07 (1.05-1.09)	4.3×10^{-14}	0.29
10p12	rs11814448	*DNAJC1*	2013	1.26 (1.18-1.35)	9.3×10^{-16}	0.02
10p15	rs2380205	*ANKRD16*	2010	0.94 (0.91-0.98)	4.6×10^{-7}	0.43
10q21	rs10995190	*ZNF365*	2010	0.86 (0.82-0.91)	5.1×10^{-15}	0.15
10q22	rs704010	*ZMIZ1*	2010	1.07 (1.03-1.11)	3×10^{-8}	0.39
10q25	rs7904519	*TCF7L2*	2013	1.06 (1.04-1.08)	3.1×10^{-8}	0.46
10q26	rs2981582	*FGFR2*	2007	1.26 (1.23-1.30)	2×10^{-76}	0.38
10q26	rs2981579	*FGFR2*	2010	1.43 (1.35-1.53)	2×10^{-76}	0.42
10q26	rs11199914	-	2013	0.95 (0.93-0.97)	1.9×10^{-8}	0.32
11p15	rs3817198	*LSP1/H19*	2007	1.07 (1.04-1.11)	1×10^{-9}	0.30
11p15	rs909116	*LSP1/H19*	2007	1.17 (1.10-1.24)	1×10^{-9}	0.30
11q13	rs614367	*CCND1/FGFs*	2010	1.15 (1.10-1.20)	3.2×10^{-15}	0.15
11q13	rs3903072	*OVOL1*	2013	0.95 (0.93-0.96)	8.6×10^{-12}	0.47
11q24	rs11820646	-	2013	0.95 (0.93-0.97)	1.1×10^{-9}	0.41
12p11	rs10771399	*PTHLH*	2011	0.79 (0.71-0.87)	4.3×10^{-35}	0.10
12p13	rs12422552	-	2013	1.05 (1.03-1.07)	3.7×10^{-8}	0.26
12q22	rs17356907	*NTN4*	2013	0.91 (0.89-0.93)	1.8×10^{-22}	0.30
12q24	rs1292011	*TBX3/MAPKAP5*	2011	0.92 (0.91-0.94)	5.9×10^{-19}	0.41
13q13	rs11571833	*BRCA2*	2013	1.26 (1.14-1.39)	4.9×10^{-8}	0.01
14q13	rs2236007	*PAX9/SLC25A21*	2013	0.93 (0.91-0.95)	1.7×10^{-13}	0.21
14q24	rs999737	*RAD51B*	2009	0.94 (0.88-0.99)	2×10^{-7}	0.24
14q24	rs8009944	*RAD51B*	2009	0.88 (0.82-0.95)	2×10^{-7}	0.24
14q24	rs2588809	*RAD51L1*	2013	1.08 (1.05-1.11)	1.4×10^{-10}	0.16
14q32	rs941764	*CCDC88C*	2013	1.06 (1.04-1.09)	3.7×10^{-10}	0.34
16q12	rs12443621	*TOX3/LOC643714*	2007	1.11 (1.08-1.14)	1×10^{-36}	0.46
16q12	rs3803662	*TOX3/LOC643714*	2010	1.20 (1.16-1.24)	1×10^{-36}	0.26
16q12	rs17817449	*MIR1972-2-FTO*	2013	0.93 (0.91-0.95)	6.4×10^{-14}	0.40
16q23	rs13329835	*CDYL2*	2013	1.08 (1.05-1.10)	2.1×10^{-16}	0.22
16q22	rs11075995	*FTO*	2013	1.07 (1.11-1.15)	4.0×10^{-8}	0.24
17q23	rs6504950	*STXBP4/COX11*	2008	0.95 (0.92-0.97)	1.4×10^{-8}	0.27
18q11	rs527616	-	2013	0.95 (0.93-0.97)	1.6×10^{-10}	0.38
18q11	rs1436904	*CHST9*	2013	0.96 (0.94-0.98)	3.2×10^{-8}	0.40
19p13	rs8170	*MERIT40*	2010	1.26 (1.17-1.35)	2.3×10^{-9}	0.18
19p13	rs2363956	*MERIT40*	2010	0.84 (0.80-0.89)	5.5×10^{-9}	0.50
19p13	rs4808801	*SSBP4/ISYNA1/ELL*	2013	0.93 (0.91-0.95)	4.6×10^{-15}	0.35
19q13	rs3760982	*KCNN4/ZNF283*	2013	1.06 (1.04-1.08)	2.1×10^{-10}	0.46
20q11	rs2284378	*RALY*	2012	1.16 (1.01-1.10)	1.1×10^{-8}	0.35
21q21	rs2823093	*NRIP1*	2011	0.94 (0.92-0.96)	1.1×10^{-10}	0.27
22q12	rs132390	*EMID1/RHBDD3*	2013	1.12 (1.07-1.18)	3.1×10^{-9}	0.04
22q13	rs6001930	*MKL1*	2013	1.12 (1.09-1.16)	8.8×10^{-19}	0.11

Ghoussaini et al. The American Journal of Pathology, 2013[28]

I.4 .2 Facteurs non-génétiques

L'agrégation de cas de cancers du sein au sein de familles a conduit à supposer l'influence d'une composante génétique pour cette maladie. Cependant, alors que la nature et la fréquence des variants génétiques est restée globalement identique au sein de la population générale depuis les 50 dernières années, l'incidence du cancer du sein n'a fait qu'augmenter. Cette augmentation serait due en partie à la modification de nos conditions de vie, exposition qualifiée d'environnementale par opposition à une exposition génétique[51].

L'incidence du cancer du sein augmente avec l'âge, est maximale aux alentours de 65 ans, puis diminue progressivement. Cependant, le risque absolu par tranche d'âge n'est pas identique en fonction de l'ethnie d'origine. En effet, les femmes africaines et afro-américaines ont un risque de cancer du sein supérieur aux femmes d'origine caucasienne jusqu'à 40 ans, alors que les tendances s'inversent entre les deux populations à partir de 40 ans[13,52].

La majorité des facteurs de risque non-génétiques sont liés au style de vie, et ont une incidence à l'échelle de la cellule aux niveaux métabolique et endocrinien[53]. Ainsi, 22% du risque de cancer du sein serait imputable à des conditions de vie dévaforables[54].

Aujourd'hui, l'influence du tabac (et notamment de la cigarette) sur le risque du cancer du sein reste encore incertaine. Bien que de nombreuses études n'aient détecté aucune association entre la consommation de cigarette et le risque de cancer du sein[55–60], d'autres ont détecté que le risque augmente avec la consommation de cigarette, et d'autant plus avec le nombre d'années de tabagisme et la précocité de l'âge de commencement[61–64]. Cette hétérogénéité des résultats entre les études peut être due à de nombreux facteurs tels que le manque de puissance statistique et la taille insuffisante des populations analysées, ou bien à cause de l'existence d'une stratification de population. C'est pourquoi une étude canadienne a analysé les variants de 36 gènes impliqués dans la métabolisation des carcinogènes dans le but de déterminer ci ceux-ci peuvent influer sur une éventuelle association entre le tabagisme et le risque de cancer du sein[65]. Leur étude inclut environ 1800 cas de cancers du sein et autant de témoins. D'après les conclusions de cette étude, aucune association n'a été détectée entre le tabagisme actif et le risque de cancer du sein, que ce soit en stratifiant ou non par statut ménopausal. Une association significative a été détectée entre le risque de cancer du sein et certains des variants étudiés situés sur les gènes *GSTT1*, *CYP2E1*, et *UGT1A7* lorsque l'on considère l'ensemble des femmes, un variant situé sur *CYP2E1* chez les femmes non-ménopausées, et un variant de *CYP1A1* chez les femmes ménopausées. D'autre part, des interactions significatives ont été détectées entre les diverses variables associées au tabagisme (statut tabagique binaire, nombre de paquets fumés par année, etc...) et des variants respectivement situés dans les gènes *CYP1A1*, *SOD2*, *CYP1B1*, *NAT1*, *UGT1A7*, certaines de ces associations étant spécifiques des femmes ménopausées ou non-ménopausées. Il semblerait donc que certains variants puissent modifier le risque de cancer du sein potentiellement induit par le tabagisme. Enfin, certaines études ayant détecté une association entre le risque de cancer du sein et le tabagisme actif se sont également intéressées au tabagisme passif. D'après ces études, il semblerait que le tabagisme passif augmente également le risque de cancer du sein et ce dans les mêmes proportions, en particulier chez les femmes non-ménopausées[66,67].

L'exposition à des radiations ionisantes est également un facteur de risque pour le cancer du sein[68,69] qui a été identifié dès les années 1980s. Plus la dose de radiations reçues est importante, plus le risque est élevé. De même, plus l'individu est jeune au moment de l'exposition, plus il a de chances de développer un cancer du sein.

La prise d'hormones exogènes, que ce soit dans le cadre de la contraception, ou bien afin de pallier aux effets secondaires de la ménopause est un facteur de risque dont les effets dépendent principalement du type d'hormones administré. La contraception orale, dont le risque est toujours sujet à controverse, n'augmenterait que très faiblement voire n'augmenterait pas le risque de cancer du sein[70]. Le traitement par remplacement d'hormones administré pour pallier aux symptômes gênants de la ménopause est quand à lui un facteur de risque avéré, notamment lorsque celui-ci est long et/ou contient de la progestérone[71]. De plus, le taux circulant d'oestradiol dans le sang est également un facteur de risque fort[72].

La densité mammaire représente la répartition de la composition du sein entre tissus adipeux (graisse) et tissus fibroglandulaires (glandes). Plus le tissu mammaire est pauvre en graisses, plus la densité mammaire est dite élevée. Une densité mammographique élevée est un facteur de risque fortement associé au cancer du sein[73]. Aujourd'hui les causes biologiques de l'association entre densité mammographique et cancer du sein sont encore mal connues, mais il semblerait que le risque de cancer dépende de la composition du tissu mammaire, mesurée par sa densité, et de sa sensibilité aux hormones telles que les oestrogènes, qui induisent la prolifération cellulaire.

Le risque de cancer du sein est aussi influencé par l'histoire reproductive, c'est à dire l'âge des premières règles, l'âge et le nombre de grossesses, la durée de la période d'allaitement[74]. Plus une femme est jeune lorsqu'elle atteint le terme de sa première grossesse, plus le risque de cancer du sein est réduit. La grossesse génère une poussée d'oestradiol et de progestérone qui engendrent le développement morphologique et fonctionnel des tissus mammaires, et tend à rendre les cellules mammaires moins susceptibles au cancer du sein. L'allaitement tend également à protéger du cancer du sein. Or, la tendance actuelle du mode de vie des femmes dans la société occidentale est d'avoir leur premier enfant plus tard, et d'allaiter pendant une faible durée.

L'influence de la surcharge pondérale sur le risque de cancer du sein a été analysé dans de nombreuses études. Parmi les mesures anthropométriques utilisées pour estimer le surpoids d'un individu, on trouve bien sûr la taille et le poids, et majoritairement l'indice de masse corporelle (IMC : rapport du poids en kg sur la taille en m élevée au carré). Les associations entre ces indices anthropométriques et le risque de cancer du sein ont été réalisées en tenant compte du statut ménopausal. Chez les femmes ménopausées, de nombreuses études ont confirmé l'association entre le surpoids/l'obésité et le risque de cancer du sein de manière robuste[75–80], le surpoids entraînant une augmentation du risque de cancer du sein (OR=1.12, intervalle de confiance à 95% : 1.12-1.16[75]). Cependant, l'association entre surpoids et risque de cancer du sein reste incertaine chez les femmes non-ménopausées[81–87]. La tendance observée serait plutôt une association inverse entre la surcharge pondérale et le risque de cancer du sein chez les femmes non-ménopausées. Le risque serait ainsi réduit de 50% chez les femmes obèses, et une augmentation de 5 $kg.m^{-2}$ de l'IMC diminuerait le risque de cancer du sein de 8%[75]. Les différences observées entre ces deux populations d'étude pourraient être liées à la différence de rôle exercé par les tissus graisseux et ainsi que leur distribution entre les femmes ménopausées et non-ménopausées. Une méta-analyse récente étudiant le rôle des facteurs anthropométriques sur le risque de cancer du sein chez les femmes non-ménopausées[88] a déterminé que l'augmentation de l'IMC d'une valeur de 5 $kg.m-2$ diminue significativement le risque de cancer du sein (OR=0.95, Intervalle de confiance à 95% : 0.94-0.97). Après stratification par ethnie, cette association reste significative avec la même tendance chez les Africains et les Caucasiens. Cependant, chez les Asiatiques, on observe une association positive entre la surcharge pondérale et le risque, le surpoids augmentant ainsi le risque de cancer du sein. L'ethnie apparaît donc comme un facteur à prendre en compte dans l'étude du risque de cancer du sein lié à la surcharge pondérale chez les femmes non-ménopausées.

La consommation d'alcool est un facteur de risque élevé. 4.5% de l'ensemble des cancers du sein sont imputables à la consommation d'alcool[89]. De même, une alimentation pauvres en

antioxydants et en folate augmente les risques de cancers du sein en diminuant les capacités de défense de l'organisme face aux attaques dirigées contre l'ADN[90,91].

I.4 .3 Interactions entre facteurs génétiques et non-génétiques

En plus des effets distincts des facteurs de prédisposition génétiques et liés à l'environnement, certains de ces facteurs peuvent interagir entre eux de manière synergique. D'autre part, certains facteurs peuvent augmenter la susceptibilité au cancer du sein uniquement en conjonction avec d'autres. Ainsi, le SNP rs17468277 localisé dans le gène *CASP8* prédispose au cancer du sein uniquement chez les consommateurs réguliers d'alcool, à hauteur de 20g d'alcool par jour, soit environ 2 verres de vin[92]. De même, une interaction a été observée entre le SNP rs3817198 et le nombre d'enfants[93,92]. Ainsi, le risque de cancer du sein chez les porteuses de ce SNP est multiplié par 1.08 (1.01-1.16) chez les femmes sans enfants, tombe à 1.03 (0.96-1.10) chez les femmes ayant un enfant, et atteint 1.26 (1.16-1.37) chez les femmes ayant eu au moins 4 enfants. Ces deux associations ont de plus été répliquées avec succès.

II. Méthodes de détection des facteurs de risque génétiques

Tous les traits phénotypiques, y compris les maladies humaines n'ont pas forcément une origine génétique. Cependant, des indices peuvent laisser penser que c'est le cas, comme l'observation de plusieurs individus atteints dans une même famille, répartis sur plusieurs générations, et ne partageant pas forcément les mêmes conditions de vie. Dans ce contexte, il devient intéressant d'envisager une origine au moins en partie génétique de la maladie, et d'explorer plus en détails le génome des individus atteints.

Notre génome est constitué d'environ 3 milliards de paires de bases, qui supportent entre 22 0000 et 25 000 gènes[94]. Identifier et localiser s'il existe un variant génomique associé à un trait phénotypique devient alors une tâche formidablement complexe. Cependant, il existe à l'heure actuelle différentes méthodes pour atteindre cet objectif. Ces méthodes reposent sur les grands principes régissant la biologie cellulaire et la génétique humaine, en particulier la génétique mendélienne, qui étudie la transmission des caractères héréditaires au niveau des individus, et la génétique des populations, qui étudie la transmission de ces caractères à l'échelle d'une population.

II.1 Notions de biologie cellulaire et de génétique des populations

L'unité structurelle de base de notre corps est la cellule. L'Homme est un organisme eucaryote. Les cellules des organismes eucaryotes sont délimitées par une membrane plasmique et possèdent un noyau baignant avec d'autres organites dans le cytoplasme (Figure 4).

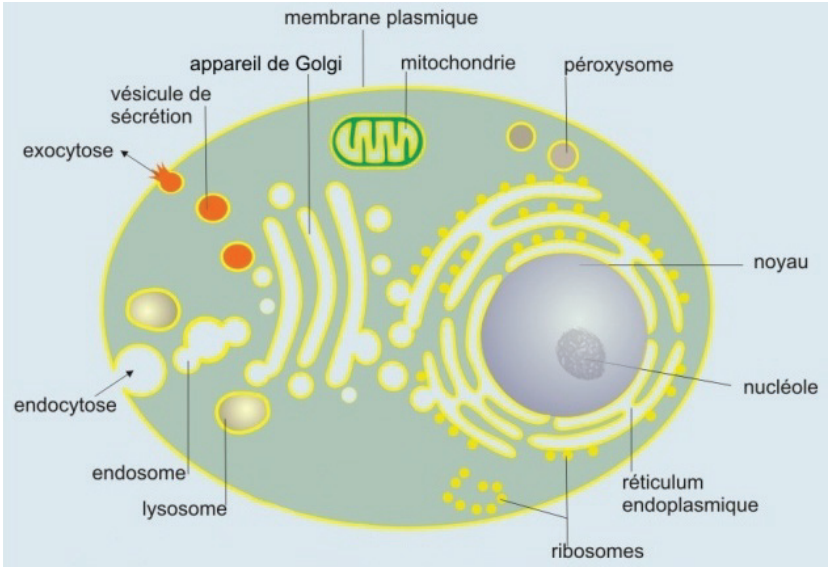
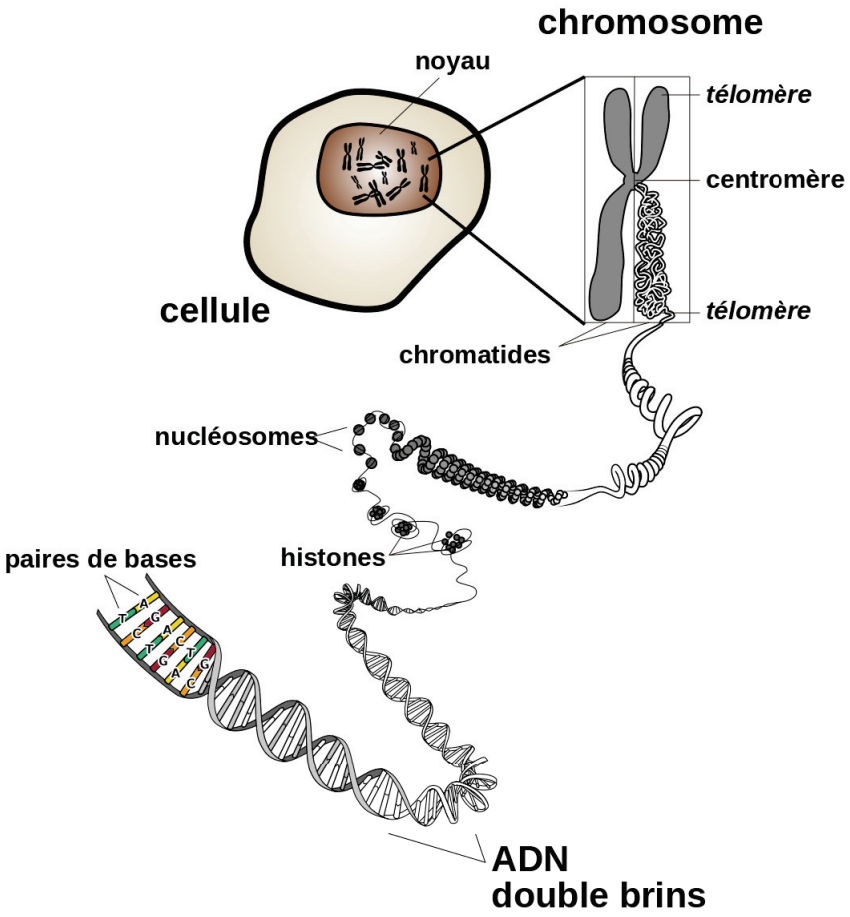


Figure 4 – Schéma d'une cellule et de ses principaux organites cellulaires

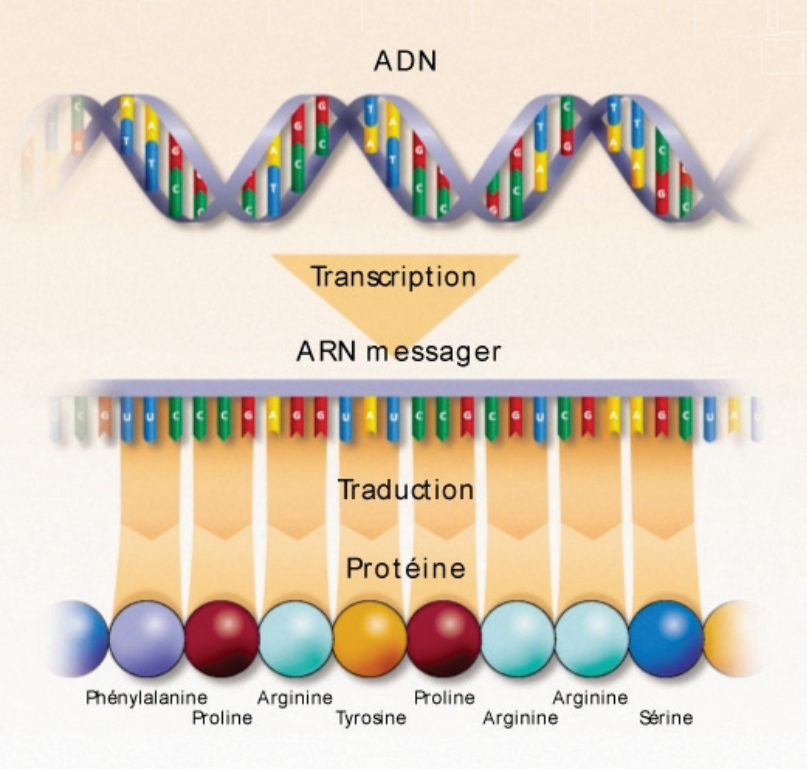
Notre corps est principalement composé de matière organique. Les principaux composants de la matière organique sont les glucides, les lipides et les protéines. Ce sont les protéines qui assurent l'immense majorité des fonctions cellulaires, en étant impliquées dans la structure de la cellule, la mobilité des composants cellulaires, la transformation chimique d'autres molécules par certaines appelées enzymes, ou encore l'accessibilité et l'expression de nos gènes. Une protéine est un enchaînement d'acides aminés. Il existe 20 acides aminés chez l'Homme dont les protéines peuvent être constituées. L'ordre dans lequel ces acides aminés s'enchaînent est défini par nos gènes. Ainsi, chaque protéine synthétisée dans nos cellules résulte de l'expression d'un de nos gènes. L'ensemble de nos gènes forment notre génome nucléaire, et sont supportés par une molécule d'ADN, et par nos chromosomes à plus large échelle. Nos gènes sont localisés dans le noyau de nos cellules (Figure 5). Notre ADN est une macromolécule ayant la forme d'une double hélice, constituée d'un enchaînement de nucléotides, des sous-unités composées d'un sucre, d'un groupement phosphate et d'une base azotée parmi les quatres existantes : l'adénine A, la cytosine C, la guanine G, et la thymine T.

Figure 5 – En fonction de l'échelle d'observation et du contexte cellulaire, notre ADN nucléaire se présente sous différents aspects.



Comme illustré de manière simplifiée sur la figure 6, afin d'aboutir à la synthèse d'une protéine, un gène est tout d'abord transcrit en un intermédiaire appelé ARN messager - ou ARNm - qui subit une série de transformations avant d'être traduit en la protéine correspondante au niveau des ribosomes, des organites cellulaires chargés de recruter et d'assembler les acides aminés dans l'ordre selon le code génétique. En effet, à chaque triplet de nucléotides, appelés aussi codons, correspond un acide aminé précis. Ainsi, une modification au niveau de la séquence en bases d'un de nos gènes peut entraîner la modification de la protéine correspondante. Cependant, pour certains acides aminés, il existe plusieurs codons correspondant à un même acide aminé, cette propriété est appelée la dégénérescence du code génétique.

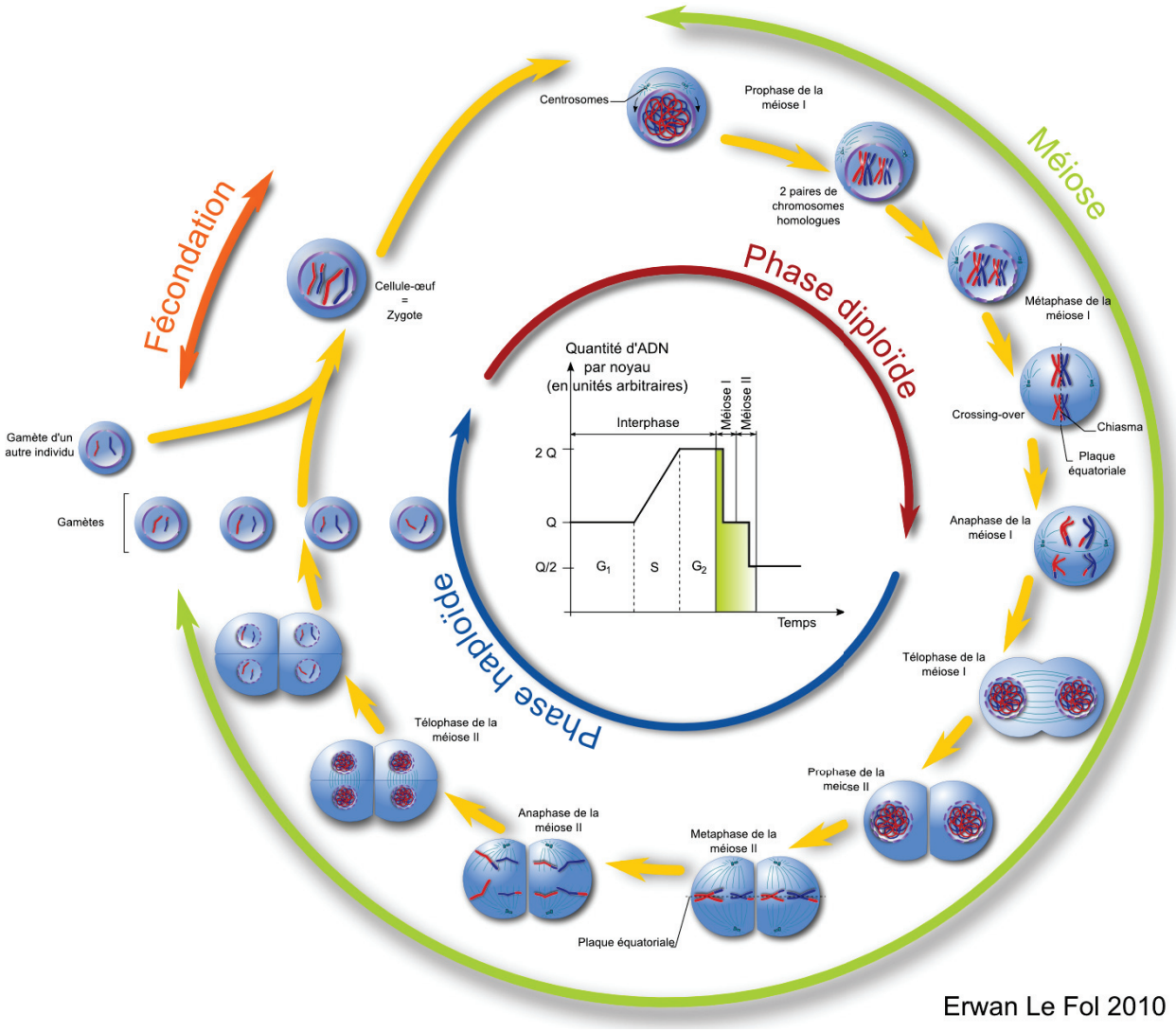
Figure 6 – Synthèse d'une protéine à partir d'un gène



Au sein de chacune des 23 paires de chromosomes composant notre génome nucléaire, l'un est d'origine paternelle, et l'autre d'origine maternelle. Chaque chromosome est formé de deux chromatides identiques accolées. Lors des divisions cellulaires successives, nos cellules alternent entre un état dans lequel les chromosomes ont une seule chromatide - état appelé *haploïde*- et un état où la seconde chromatide est présente à la suite de la réplication à l'identique de la première - état appelé *diploïde*.

Les cellules reproductrices humaines, appelées gamètes, sont le produit d'une division cellulaire spécifique appelée méiose, qui à partir d'une cellule diploïde, produit quatre gamètes (Fig. 7). Lorsqu'une cellule n'est pas en cours de division, l'ADN baigne dans le cytoplasme sous sa forme filaire et condensée, lui permettant ainsi d'interagir avec diverses protéines nécessaires à l'expression et à la régulation génique. Cependant, lorsqu'une cellule entre en division cellulaire,

Figure 7 – Représentation des différentes étapes successives de la méiose



que ce soit en mitose - division cellulaire classique lors de laquelle une cellule mère se divise en deux cellules-filles - ou en méiose, la chromatine se condense, donnant ainsi aux chromosomes leur structure à simple ou double chromatide. Durant cette division cellulaire particulière qu'est la méiose, les paires de chromosomes homologues s'alignent sur l'axe médian de la cellule (étape Métaphase de ma méiose I) avant que chacun des chromosomes homologues d'une paire migre à un des pôles de la cellule (étape Anaphase de la méiose I). Lorsqu'ils sont ainsi alignés, un évènement essentiel au brassage génétique peut avoir lieu : la recombinaison méiotique.

Lors d'un évènement de recombinaison, schématiquement représenté sur la figure 8, les chromosomes homologues de chaque paire sont très proches physiquement, les bras des chromosomes peuvent alors former des enjambements - ou *crossing-over* - et des portions des deux chromosomes homologues peuvent être échangées. À l'issue de la recombinaison, les deux chromosomes homologues sont donc légèrement différents des deux chromosomes initiaux, et portent des

combinaisons d'allèles potentiellement différentes des combinaisons portées initialement par les chromosomes homologues. Ainsi, dans l'exemple présenté sur la figure 8, les combinaisons d'allèles parentaux, également appelées haplotypes, étaient *ABC* et *abc*. Après recombinaison, deux des quatre chromatides ont conservé ces haplotypes mais les deux autres portent maintenant les haplotypes *ABc* et *abC*. Ce brassage allélique interchromosomique est une des principales sources de variabilité génétique et phénotypique existantes.

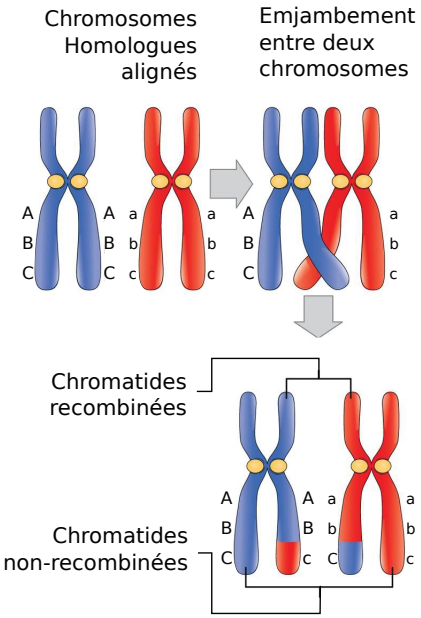


Figure 8 – Évènement de recombinaison : schématisation d'un enjambement entre chromosomes homologues

On distingue, pour deux loci situé sur un même chromosome, la distance physique et la distance génétique. La distance physique correspond au nombre de paires de bases séparant ces deux loci. La distance génétique mesure la fréquence à laquelle les allèles de ces deux loci recombinent entre eux. Cette fréquence de recombinaison se mesure en centimorgan, noté cM. 1 cM représente 1% de recombinaison, soit 1 enjambement en moyenne sur la distance séparant ces deux loci pour 100 méioses. Ce taux de recombinaison varie entre 0, ce qui implique que les loci considérés sont totalement liés, à 0.5 pour des loci génétiquement indépendants. Sans que la relation soit parfaitement linéaire, plus la distance physique (mesurée en paires de bases) entre deux gènes situés sur un même chromosome est élevée, plus le taux de recombinaison associée le sera également.

Les positions génomiques au niveau desquelles des évènements de recombinaison ont lieu sont appelés points de cassures. La répartition des points de cassures le long des chromosomes n'est pas homogène chez l'Homme[95]. Dans certaines régions, dont la longueur est de l'ordre du milliers de paires de bases, les évènements de recombinaison sont jusqu'à 1000 fois plus fréquents que dans les régions avoisinantes. Les mécanismes de régulation de la recombinaison du génome humain sont encore très peu connus. Cependant, le gène *PRDM9* situé sur le chromosome 17, a récemment été identifié comme un des acteurs majeurs de la recombinaison chez l'Homme[96].

Ce gène code pour la protéine PRDM9 possédant un domaine appelé domaine à doigts de zinc - ou *Zinc finger domain containing protein* - ce qui lui confère la possibilité de se lier sur l'ADN. De plus, elle est également impliquée dans la réparation des cassures double-brins, mécanisme nécessaire à la recombinaison méiotique. Des prédictions *in silico* ont révélé que cette protéine pourrait se lier sur des motifs de l'ADN dont les séquences des points chauds de recombinaison sont enrichis. Les variations au sein de *PRDM9* ont été corrélées avec des modifications de la fraction des cross-overs ayant lieu au niveau de points chauds de recombinaison, avec des différences d'utilisation de ces points chauds, affectant la distribution des évènements de recombinaison au niveau du génome entier[97].

Les évènements de recombinaison n'étant pas distribués de manière homogène le long des chromosomes, les allèles situés entre deux points chauds de recombinaison ont une probabilité élevée d'être transmis ensemble à la descendance. Ces allèles étant plus fréquemment cotransmis à la descendance que ne le supposerait le hasard, on parle alors de déséquilibre de liaison, également appelé déséquilibre gamétique - ou *Linkage Desequilibrium*, souvent noté LD. On dit que deux allèles de deux loci distincts sont en déséquilibre de liaison si la probabilité d'observer ces deux allèles chez un même individu est plus élevée que ce que le hasard supposerait, c'est à dire si cette probabilité est supérieure au produit des fréquences alléliques respectives de ces deux allèles dans une population donnée. Les régions séparées par des points chauds de recombinaison sont appelés blocs de déséquilibre de liaison. À l'échelle d'une population, le déséquilibre de liaison entre deux allèles se mesure à l'aide d'indicateurs dont les plus utilisés sont : D, D' , et r^2.

Soient α et β deux loci bialléliques, dont les allèles sont respectivement $\{A; a\}$ et $\{B; b\}$. La fréquence de l'allèle 1 de ces loci est respectivement notée p_A et p_B. La fréquence de l'haplotype constitué par les allèles A et B est notée p_{AB}. On appelle D la mesure de la déviation entre le fréquence haplotypique observée p_{AB} et le produit des fréquences alléliques p_A et p_B :

$$D = p_{AB} - p_A.p_B$$

La gamme de valeurs que peut prendre D dépend des fréquences alléliques des allèles considérés. Il est donc parfois plus facile de manipuler D' qui correspond à D, mais normalisé par sa valeur théorique maximale en fonction des fréquences alléliques :

$$D' = \begin{cases} \frac{D}{min\{\ p_A.p_a,\ p_B.p_b\ \}} & \text{if } D > 0 \\ \frac{D}{min\{\ p_A.p_B,\ p_a.p_b\ \}} & \text{if } D < 0 \end{cases}$$

D' est alors compris entre -1 et 1. Lorsque $|D'|$ prend une valeur proche de 1, alors les allèles sont en déséquilibre de liaison. Lorsque les fréquences alléliques des deux allèles considérés sont comparables, alors une valeur élevée de D' indique qu'un des allèles peut être utilisé

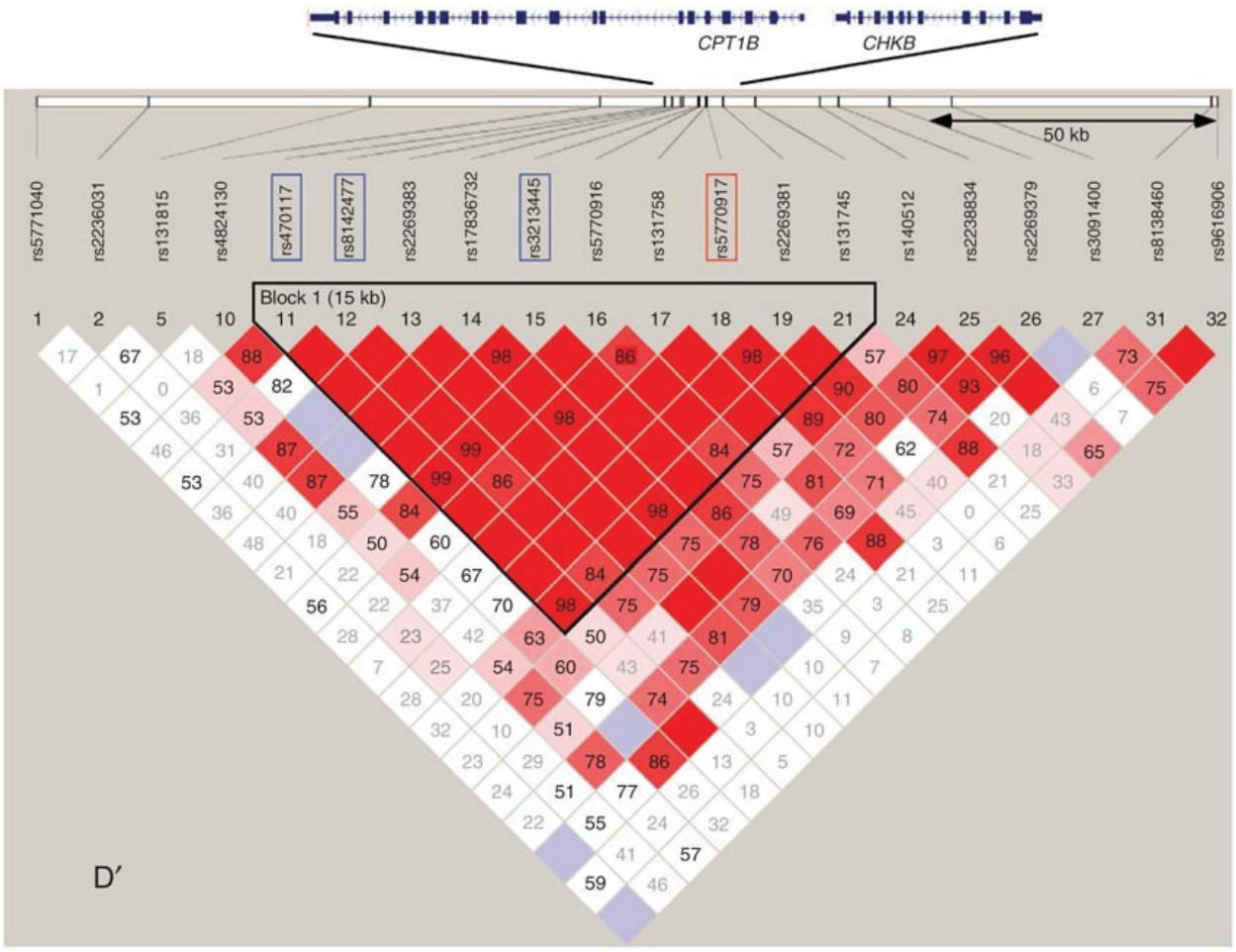
comme substitut pour l'autre. Un troisième indicateur est couramment utilisé : le coefficient de corrélation au carré entre les allèles des deux loci considérés :

$$r^2 = \frac{D^2}{p_A.p_a.p_B.p_b}$$

r^2 varie également entre 0 et 1. Lorsque r^2 est proche de 1, les allèles sont en déséquilibre de liaison. Une valeur de r^2 élevée indique qu'un allèle peut être utilisé comme substitut de l'autre. Le seuil communément utilisé est 0.8 ; si $r^2 > 0.8$, alors on estime qu'on peut légitimement utiliser le génotype d'un des loci pour inférer celui du second.

D' est l'indicateur le plus couramment utilisé pour représenter les blocs de déséquilibre de liaison. La figure 9 représente le déséquilibre de liaison entre les marqueurs situés à proximité des gènes *CPT1B* et *CHKB* sur le chromosome 22. Un polymorphisme situé entre ces deux gènes a été observé associé à une hausse de la susceptibilité à la narcolepsie[98]. Sur ce type de graphe, le déséquilibre de liaison de tous les couples de marqueurs inclus est indiqué. Le code couleur est le suivant : en blanc, absence de déséquilibre de liaison ; en rose dégradé et en bleu, déséquilibre de liaison intermédiaire ; en rouge, fort déséquilibre de liaison. Dans chaque case est indiquée la valeur de D' pour le couple de marqueurs correspondant. La valeur est omise lorsque $D' = 1$. Cette représentation graphique révèle de manière explicite l'existence d'un bloc de déséquilibre de liaison. En effet, on observe un ensemble de 10 marqueurs, positionnés de manière consécutive sur le génome, qui sont tous en déséquilibre de liaison très fort les uns avec les autres.

Figure 9 – Représentation graphique du déséquilibre de liaison



La région représentée se situe au niveau du chromosome 22, entre les positions 50 902 390 et 51 104 686 (Version du génome : *hg*19), dans laquelle se situent les gènes *CPT1B* et *CHKB*. Les valeurs indiquées représentent D' pour tous les couples de marqueurs indiqués. Par simplification, lorsque $D' = 1$, la valeur n'est pas indiquée sur le graphe. *D'après Miyagawa et. al., Nature Genetics, 2008* [98]

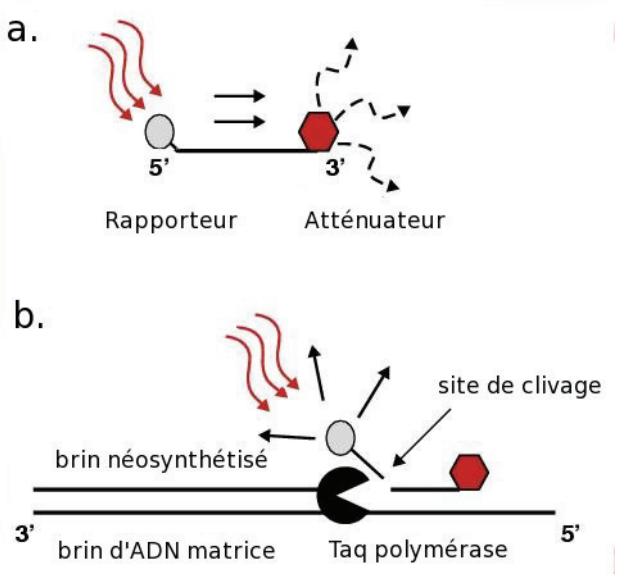
II.2 Techniques d'acquisition de données génétiques

Les deux principales technologies d'acquisition de données génétiques évoquées ici sont le génotypage et le séquençage. Ces deux technologies nécessitent d'avoir à disposition un échantillon biologique à partir duquel il est possible d'extraire l'ADN. Pour l'étude du génome constitutionnel, l'échantillon biologique le plus couramment utilisé est le sang, et l'ADN est généralement extrait à partir des lymphocytes, et amplifié.

II.2 .1 Le génotypage

Le génotypage est l'ensemble des méthodes permettant d'établir le génotype en certains endroits précis du génome, en particulier au niveau des polymorphismes d'une seule paire de bases. On utilise pour cela des puces à ADN. Il existe plusieurs techniques différentes de génotypage, mais toutes utilisent le principe de complémentarité entre les fragments d'ADN étudiés et des sondes ciblant les polymorphismes d'intérêt, ainsi que l'émission de fluorescence pour identifier quel allèle se trouve à la position étudiée. Le principe de la technique TaqMan® est présentée dans la figure 10.

Figure 10 – Principe de la technique TaqMan®



Les sondes employées (Figure 10a) sont longues de quelques dizaines de nucléotides et sont complémentaires de la région où se situe le SNP qu'elles ciblent. Plus précisément, les sondes employées fonctionnent par couples, dont les deux membres sont identiques à l'exception du nucléotide correspondant au SNP étudié, chacun des deux sondes portant un des deux allèles du polymorphisme. De plus, elles portent en leur extrémité 5' un rapporteur émettant un rayonnement par fluorescence, et en 3' un atténuateur, ayant la capacité d'absorber le rayonnement émis par le rapporteur s'il se situe à une faible distance. Les rapporteurs correspondant aux sondes associées à chacun des allèles d'un SNP donné sont couplés à des fluorochores différents. L'enzyme Taq est une polymérase exonucléase, elle peut donc incorporer des nucléotides et procéder à l'élongation du brin néosynthétisé, tout en exerçant sa fonction d'exonucléase à l'autre

extrémité (Figure 10b). Lorsque l'élongation atteint la sonde, si l'allèle porté par la sonde est complémentaire du brin d'ADN et s'est donc effectivement hybridé sur ce brin, la Taq peut alors cliver le nucléotide correspondant, ce qui relargue le rapporteur. Une fois le rapporteur relargué et éloigné de l'atténuateur, sa fluorescence peut être captée. Si l'allèle de la sonde n'est pas complémentaire de la séquence d'ADN, la Taq ne peut effectuer de clivage, et le rapporteur n'est pas libéré de l'atténuateur. La détection des pics de fluorescence et de leur longueur d'onde permet de savoir quelle sonde est complémentaire du brin d'ADN étudié, et donc quel allèle est présent au niveau du polymorphisme.

Les technologies de génotypage sont relativement simples, à la fois conceptuellement et dans leur mise en place. Cependant, elles sont soumises à des biais qui leur sont inhérents. Par exemple, il peut y avoir hybridation croisée ; certaines sondes ne s'apparient pas avec leur ADN cible, mais avec un ADN dont la séquence est proche. L'hybridation des sondes, et la force avec laquelle elles s'hybrident sont en partie dépendantes de la séquence de la sonde. En effet, la formation d'une liaison entre deux bases complémentaires G et C est plus forte que la liaison formée entre l'autre couple de bases complémentaires A et T. Cette différence de force d'hybridation peut également contribuer à l'apparition d'hybridations croisées.

II.2 .2 Le séquençage

Le séquençage d'ADN est une méthode d'acquisition de données génomiques qui ne vise pas à connaître le génotype à une position précise, mais cherche à extraire le génotype de l'ensemble des positions des fragments d'ADN analysés. Il est possible de séquencer l'ensemble de l'ADN du génome, on parle alors de séquençage de génome complet. Il est également fréquent de ne séquencer que l'exome, soit les portions du génome codant pour des protéines. Enfin, il est tout à fait possible de ne capturer que certaines régions d'intérêt et de les séquencer. De même que pour le génotypage, l'ADN est extrait, amplifié parfois, et fragmenté.

Il existe de nombreuses technologies de séquençage. Le séquençage Sanger est la première techniques de séquençage a avoir été utilisée dans les années 1970. L'échantillon d'ADN a séquencer est fragmenté en fragments de taille allant jusqu'à 3 kb maximum. Chaque fragment est alors amplifié par PCR de manière indépendante. Après amplification, le fragment amplifié est mis en présence d'un mélange des 4 désoxyribonucléotides constituant notre ADN (dATP, dTTP, dCTP, dGTP, et dont le mélange est noté dNTP). On ajoute également une faible proportion de ddNTPs, soit des didesoxyribonucléotides : ce sont des dNTPs privés de 2 groupements $-OH$. L'absence de ces groupements empêche toute élongation du brin d'ADN après incorporation d'un ddNTP. Les ddNTPs introduits sont de plus marqués par un fluorochrome spécifique de chaque base A, C, G, T. Enfin, des amorces et des ADN-polymérases sont également introduites dans le mélange réactionnel. Ainsi, les amorces s'hybrident sur les brins d'ADN amplifiés, et l'ADN-polymérase va alors synthétiser le brin d'ADN complémentaire de ce fragment en incorporant les nucléotides disponibles. Cependant, lorsqu'un ddNTP est incorporé, l'élongation cesse (Figure 11). Ainsi, l'incorporation des ddNTPs qui a lieu de manière stochastique aboutit à un ensemble de fragments de taille variable, et dont le dernier nucléotide incorporé est marqué (Figure 12). On fait ensuite migrer les fragments sur gel par électrophorèse, ils vont alors s'aligner en fonction de leur taille. Enfin, un laser et un détecteur

sont ensuite utilisés afin de déterminer quelle base se situe à chaque position de la séquence.

Figure 11 – Séquençage Sanger : Polymérisation

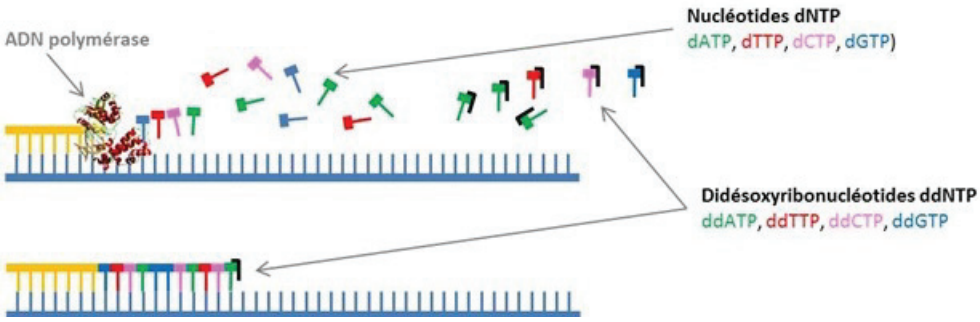
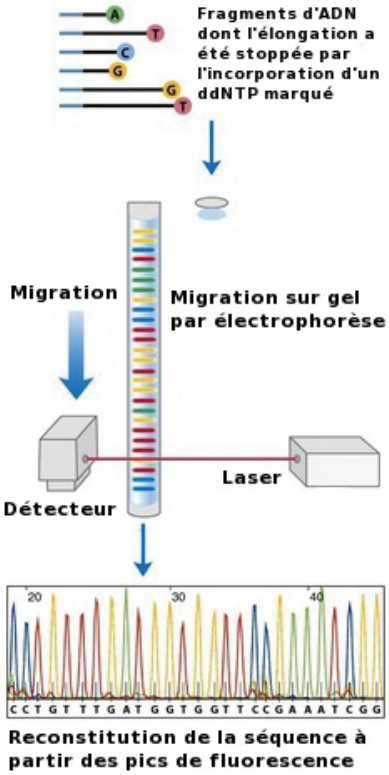


Figure 12 – Séquençage Sanger : Détermination de la séquence à partir des fragments néosynthétisés



Le séquençage Sanger est très fiable et reste la technique de séquençage de référence, présentant très peu de faux-positifs. Cependant, cette méthode reste relativement lente, puisque l'amplification et la PCR de séquençage et la migration sur gel doivent être faites de manière indépendante pour chaque fragment. C'est pourquoi de nouvelles technologies de séquençage appelées technologies à haut débit ou encore NGS pour *Next-Generation Sequencing* ont fait

leur apparition. Ces technologies, indépendemment du principe de séquençage qu'elles utilisent, ont la propriété d'être massivement parallélisées, et de pouvoir séquencer un nombre élevé de fragments d'ADN de manière simultanée. Elles sont donc beaucoup plus efficace en nombre de paires de bases séquencées par unité de temps. Cependant, la sensibilité et la spécificité de ces nouvelles technologies de séquençage n'atteignent pas encore celles de la méthode Sanger. Les technologies Illumina et Ion Torrent en sont deux exemples.

La technologie d'Illumina est un type de séquençage par synthèse. Une amorce vient s'apparier sur le fragment d'ADN à séquencer. On introduit également des enzymes polymérases pouvant procéder à l'élongation de l'amorce par complémentarité avec le fragment d'ADN sur lequel l'amorce est hybridée. On introduit alors des nucléotides libres dans le milieu, chaque type de nucléotide A, C, T, ou G étant couplé à un fluorochore différent. Ces nucléotides sont d'autre part munis d'un groupement protecteur qui empêche la polymérase d'incorporer un second nucléotide à la suite du premier. À chaque cycle d'incorporation d'un nucléotide, la fluorescence est mesurée, et en fonction de la longueur d'onde émise, on sait quelle base à été incorporée. On supprime ensuite le groupement protecteur, et un nouveau cycle recommence. À la fin du processus, on a reconstitué la séquence de l'ensemble des fragments d'ADN analysés. Ces fragments sont appelés lectures, ou *reads*.

Ion torrent utilise une autre technologie de séquençage. C'est également une méthode de séquençage par synthèse, mais basée sur la détection de variations de pH. En effet, l'incorporation d'un nucléotide par la polymérase rejette un ion H^+, ce qui fait varier le pH du milieu. De la même manière que la technologie Illumina, des nucléotides sont injectés de manière successive dans le milieu réactionnel par cycle, et à chaque injection d'un type de nucléotide, une détection de la variation du pH est effectuée. Si le pH varie, il y a eu élongation. Contrairement à la technologie d'Illumina, les nucléotides injectés ne possèdent pas de groupement protecteur, la polymérase peut donc incorporer plusieurs nucléotides identiques à la suite. Ainsi, si la séquences du brin d'ADN séquencé contient plusieurs bases A d'affilée, alors autant de T seront incorporés dans le brin néosynthétisé. Le nombre de nucléotides ainsi incorporés est déterminé par la variation de pH détectée qui augmente avec le nombre d'incorporations.

L'analyse des données de séquençage nécessite un traitement bioinformatique beaucoup plus important que celui des données de génotypage. De plus chaque technologie de séquençage présente ses propres limites et biais. Le séquençage est à l'heure actuelle encore beaucoup plus coûteux que le génotypage.

II.3 Analyse de liaison génétique, ou *Linkage Analysis*

Le premier type d'analyse généralement effectué dans l'identification d'une éventuelle composante génétique dans l'étude d'une maladie est l'analyse de liaison génétique. Ce type d'analyse permet en effet de localiser grossièrement au niveau d'une région chromosomique, à quel endroit se situe l'élément génétique associé au trait phénotypique[99]. Les analyses de liaison génétique sont basées sur l'analyse au sein d'une famille des génotypes dont certains sont issus d'évènements de recombinaison.

II.3 .1 Principe des analyses de liaison génétique

Pour effectuer une analyse de liaison génétique, on génotype un ensemble des marqueurs au sein des individus d'une même famille dans laquelle certains sont affectés et d'autres pas. L'étude de la transmission des allèles de marqueurs étudiés au sein de cette famille, en tenant compte de l'existence du mécanisme de recombinaison lors de la méiose, permet alors d'identifier dans quelle région chromosomique le variant à l'origine de la maladie se situe. L'objectif est de déterminer pour chaque marqueur étudié si le variant causal est proche de lui en terme de distance génétique. Plus on a d'individus recombinants et de familles dont le pédigrée est connu, plus la région identifiée sera de taille limitée, et la localisation précise. L'outil statistique utilisé dans ce type d'analyse est le LODscore, pour *Logarithm of the odds score*. C'est le logarithme du rapport des vraisemblances entre l'hypothèse proposée (liaison génétique entre le marqueur et l'élément causal) et l'hypothèse contraire (absence de liaison).

On cherche à localiser la position d'un marqueur génétique biallélique à l'origine d'une maladie donnée. Soit M un marqueur dont 6 allèles sont observés au sein d'une famille, notés $A1$ à $A6$. Soit θ le taux de recombinaison entre le marqueur génétique à identifier et M. La figure 13 représente l'arbre généalogique d'une famille dont les individus sont répartis sur 3 générations. La maladie génétique étudiée est autosomale - il y autant d'individus des deux sexes atteints, le gène impliqué n'est donc pas porté par un des chromosomes sexuels, mais par un chromosome autosomique - et dominante - un individu peut être atteint en ne portant qu'un seul allèle causal. De plus, cette maladie est supposée être à pénétrance complète, c'est à dire qu'un individu porteur de l'allèle causal à 100% de chances de présenter le phénotype pathologique correspondant.

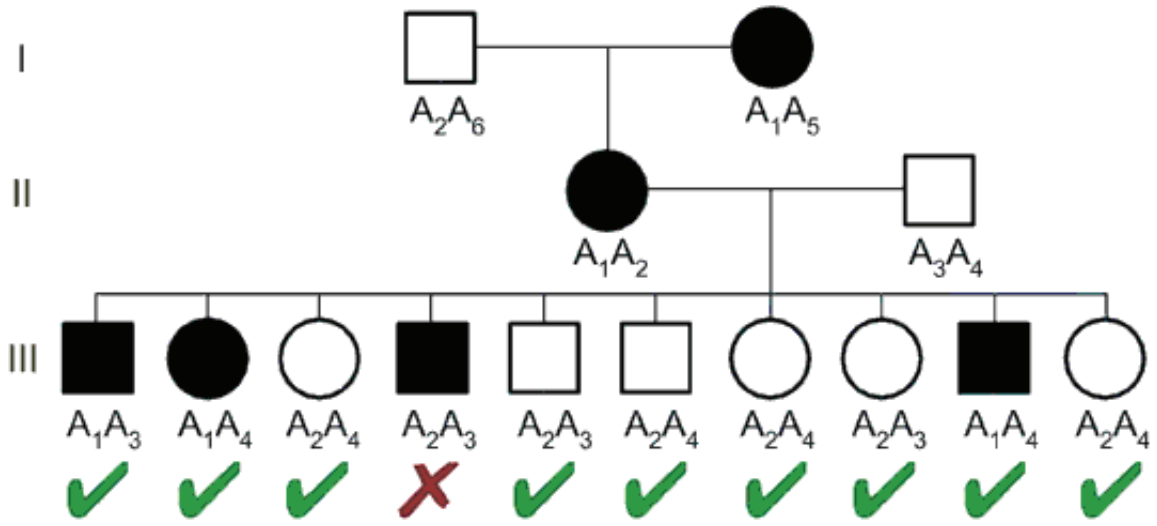


Figure 13 – Exemple de pedigree

Les individus de cette famille, répartis sur 3 générations - I, II, et III - sont représentés selon leur sexe - disque pour les femmes, carré pour les hommes - et leur statut - forme pleine pour un individu atteint, vide pour un non atteint. Le génotype biallélique de chaque individu est indiqué, avec des allèles allant de $A1$ à $A6$. Le quatrième individu de la génération III présente un génotype recombinant, indiqué par une croix rouge.

Sur cette figure, on peut observer que tous les individus atteints portent l'allèle A1, excepté l'individu III-4, qui porte le génotype A2A3. Cet individu est très vraisemblablement recombinant. En effet, on peut formuler l'hypothèse que l'élément causal et l'allèle A1 sont généralement transmis ensemble, mais que lors de la méiose ayant abouti à la synthèse du gamète à partir duquel l'individu III-4 s'est développé, un évènement de recombinaison situé entre A1 et ce variant a conduit à séparer les régions chromosomiques portant ces deux allèles. L'individu III-4 n'a donc pas hérité de l'allèle A1, mais de l'élément causal de la maladie. Sur les $n = 10$ individus de la génération III, on observe donc $r = 1$ recombinant. Ce qui conduit à une estimation du taux de recombinaison $\theta = \frac{r}{n} = \frac{1}{10} = 0.1$. On va alors tester statistiquement les hypothèses suivantes :

H_0 : *Il n'y a pas de liaison génétique entre M et le marqueur recherché*
$\theta = \frac{1}{2}$

H_1 : *Le marqueur recherché et M sont liés génétiquement avec un distance génétique de* θ
$\theta \neq \frac{1}{2}$

On calcule pour cela la statistique du LODscore $Z(\theta)$:

$$Z(\theta) = LOG\left(\frac{L(\theta)}{L(\theta = \frac{1}{2})}\right) \qquad \text{avec :} \quad L(\theta) = \left(\frac{\theta}{2}\right)^r . \left(\frac{1-\theta}{2}\right)^{n-r}$$

À partir de la valeur estimée de θ basée sur le nombre d'enfants total et recombinants, et en appliquant cette formule, on obtient une valeur numérique pour $Z(\theta)$. Si, pour diverses raisons, on ne dispose pas d'une estimation de θ, on prend pour $Z(\theta)$ la valeur correspondant au maximum de vraisemblance de $Z(\theta)$ en fonction de toutes les valeurs possibles de θ. Dans l'exemple présenté, $\theta = \frac{1}{10}$, ce qui donne $Z(\theta = \frac{1}{10}) = 1.598$. On conclut sur la probabilité de l'existence d'une liaison génétique entre le trait associé à la maladie et le marqueur M en comparant la valeur de $Z(\theta)$ avec des seuils établis :

- ◇ Si $Z(\theta) < -2$: il est 100 fois plus probable qu'il n'y ait pas de liaison génétique entre les deux marqueurs étudiés plutôt qu'ils soient liés. On conclut qu'il n'y a statistiquement pas de liaison génétique entre les marqueurs.
- ◇ Si $-2 < Z(\theta) < 3$: zone intermédiaire dans laquelle on ne peut pas conclure sur l'existence ou l'absence d'une liaison génétique. On ne rejette pas H_0.
- ◇ Si $Z(\theta) > 3$: il est 1000 fois plus probable que les marqueurs soient génétiquement liés plutôt qu'ils ne le soient pas. On rejette H_0, et on accepte H_1. On conclut que les marqueurs étudiés sont génétiquement liés, et ce de manière statistiquement significative.

Dans le cas de l'exemple ci-dessus, $Z(\theta) = 1.598$, et $-2 < Z(\theta) < 3$: on ne peut pas conclure de manière statistiquement significative sur l'existence d'une liaison génétique entre M et l'élément génétique recherché. Cependant, cette valeur nous renseigne sur la tendance observée, puisque qu'une valeur de 1.598 signifie qu'il est $10^{1.598} = 40$ fois plus probable que ces deux marqueurs soient liés plutôt qu'ils ne le soient pas.

Dans le cas où deux marqueurs génétiques sont liés, il existe des méthodes de calcul, appelées fonctions de cartographie, permettant d'estimer la distance génétique séparant ces deux loci

en fonction des taux de recombinaison, qui eux ne sont pas additifs. Une des fonctions de cartographie les plus simples est la fonction de Haldane, qui suppose que les évènements de recombinaison se produisent de manière aléatoire et indépendamment les uns des autres. Cette fonction permet d'estimer la distance génétique d séparant les marqueurs en fonction du taux de recombinaison θ :

$$d = -\frac{1}{2}.ln(1 - 2\theta)$$

On sait cependant que les hypothèses sur lesquelles repose la fonction de Haldane ne sont en réalité pas toutes vérifiées. En effet, la probabilité d'observer un évènement de recombinaison n'est pas identique en tout point du génome, puisque des points chauds de recombinaison ont été mis en évidence. De plus, ces évènements ne sont pas parfaitement indépendants les uns des autres, puisque mécanistiquement, lorsqu'un évènement de recombinaison a lieu, il est très peu probable qu'un second ait lieu dans sa proximité immédiate, à cause des contraintes physiques et d'encombrement spatial exercées (phénomène appelé interférence). D'autres fonctions de cartographies plus complexes, telles que la fonction de Kosambi, ont donc été mises au point afin de tenir compte de l'ensemble de ces paramètres pour estimer de manière plus précise la distance génétique séparant les marqueurs étudiés.

II.3 .2 Avantages et inconvénients

Les analyses de liaison génétique sont un outil optimal afin d'identifier les causes génétiques de maladies monoalléliques. En effet, si la maladie est relativement fréquente, il sera possible d'étudier de nombreuses familles, et d'estimer précisément où se situe le marqueur génétique causal. La plupart des maladies monoalléliques sont rares. Cependant, les gènes impliqués ont parfois pu être identifiés à la suite de l'analyse de liaison d'une seule famille, comportant des malades sur plusieurs générations.

Cependant, la maladie étudiée peut être rare, et peu de familles sont disponibles pour mettre en place des analyses de liaison. D'autre part, certaines maladies sont dites à pénétrance incomplète : les individus portant la mutation pathogène n'expriment pas forcément le phénotype de la maladie, ils ne sont donc pas encore diagnostiqués au moment où les études de liaison génétique sont effectuées. Leur statut affecté ou non affecté étant incertain, les analyses s'en trouvent complexifiées. Le modèle d'expression génique de la maladie peut également être bien plus complexe qu'une simple dominance allélique. Un modèle récessif nécessite que le porteur ait deux allèles défectueux, alors que dans un modèle à codominance, les deux allèles s'expriment. Pour ce dernier modèle, en fonction des allèles portés, les individus atteints peuvent présenter divers degrés d'expression de phénotype de la maladie. Ces types de modèles nécessitent d'adapter les méthodes de calcul pour les analyses de liaison génétique. Enfin, ce type d'analyse n'est pas optimal pour étudier des maladies multifactorielles, impliquant plusieurs gènes ou marqueurs génétiques.

Des analyses de liaison génétique dans les familles atteintes de la forme familliale du cancer du sein ont été mises en place et ont tout d'abord permis d'identifier une région d'intérêt

sur le chromosome 17 au niveau de la portion 17q21[100] en 1990. En 1994, un premier gène de susceptibilité au cancer du sein, *BRCA1* - pour *BReast Cancer susceptibility gene 1* - a été identifié[101]. La même année, une autre région située sur le chromosome 13 a été localisée[102]. L'année suivante, le gène *BRCA2* situé sur cette région a été formellement identifié et précisément localisé[103].

II.4 Études d'association

II.4 .1 Principe

Une étude d'association génétique est une étude ayant pour but de détecter une association entre un facteur d'exposition génétique et la survenue d'un évènement, et se base généralement sur des données de génotypage. À la différence des études de liaison génétique qui sont réalisées au niveau d'un nombre restreint d'individus apparentés, les études d'association sont effectuées au niveau populationnel, avec un nombre conséquent d'individus, apparentés ou non. Bien qu'il existe plusieurs types d'études permettant de faire cela - notamment les études de cohortes - nous nous focaliserons sur les études de type cas-témoins, de loin les plus fréquentes en épidémiologie analytique. Des polymorphismes d'une seule paire de bases sont génotypés à la fois sur des individus présentant un trait phénotypique d'intérêt, et sur d'autres ne le présentant pas. Ces individus sont communément appelés des cas (ou individus affectés) et des témoins (individus non-affectés), et ce même lorsque l'évènement étudié est autre que la survenue d'une maladie. On cherche ensuite à détecter parmi les SNPs génotypés si certains ont une fréquence différente dans le groupe des cas et dans le groupe des témoins, c'est à dire si certains SNPs sont plus ou moins souvent observés dans un groupe que dans l'autre. Si un SNP est plus souvent observé dans le groupe des individus présentant la caractéristique d'intérêt, on dit que ce SNP est associé avec ce trait phénotypique. Dans le cas contraire, lorsque l'allèle SNP est moins fréquent chez les individus qui présentent ce trait, on dit que le SNP est inversement associé avec ce trait.

Une étude cas/témoin est une étude observationnelle - on ne contrôle pas l'exposition des cas, contrairement à des études telles que celles utilisées dans les essais cliniques - et rétrospective - l'étude commence alors que les cas sont déjà déclarés, et on analyse leur exposition avant le diagnostic uniquement après leur inclusion dans l'étude. Une étude cas/témoin peut être équilibrée - c'est à dire inclure un nombre égal de cas et de témoins - ou déséquilibrée - en incluant généralement plus de témoins que de cas. Les témoins peuvent être appariés aux cas sur des critères tels que l'âge, le sexe, l'ethnie, ou non. De manière générale, une étude cas/témoin est mise en place lorsque l'on cherche à savoir si un facteur d'exposition est associé à la survenue d'un évènement ou la manifestation d'une caractéristique. Dans le cadre d'une analyse d'association où l'on étudie l'influence des polymorphismes sur le risque de développer une maladie, l'exposition étudiée est le fait de porter l'un au l'autre allèle, généralement l'allèle le plus rare, appelé également allèle alternatif.

Les études cas-témoins présentent certains avantages : elles sont rapides à mettre en place, et peu coûteuses. Cependant, elles peuvent présenter plusieurs types de biais qu'il convient de connaître, d'anticiper, ou de contrôler. Ces biais sont de trois types : les biais de sélection, les biais de classement, et les biais de confusion. Il y a biais de sélection lorsque les sujets choisis pour constituer l'échantillon analysé ne sont pas représentatifs de la population source ou ne

sont pas adaptés à la question posée. L'existence d'un biais de sélection est souvent associé à une définition imprécise, incomplète ou inadaptée des critères d'inclusion de l'étude. Par exemple, le biais de survie a pour conséquence d'enrichir l'échantillon étudié en individu atteints, mais dont la sévérité de la maladie est plus faible que dans la réalité, puisque les cas les plus graves sont décédés.

Un biais de classement affecte la manière dont l'exposition et la variable de sortie sont mesurées chez les individus inclus dans l'étude. Ce type de biais est souvent associé à la méthode de recueil ou à la nature des informations. Par exemple, un biais de mémoire est souvent observé dans le cadre d'études cas-témoin rétrospectives, dans lesquelles on demande aux individus inclus, souvent par questionnaire, d'estimer eux-mêmes leur niveau d'exposition à un facteur externe (alimentation, comportements, antécédents familiaux, etc...). Les individus, bien que de bonne foi, ne sont pas objectifs du fait de leur inclusion dans l'étude, et fournissent généralement des informations parfois exagérées ou incomplètes sur la mesure de leur exposition. Dans les études d'association où le facteur d'exposition est génétique, le biais de mémoire n'intervient pas, puisque la mesure de l'exposition est établie par génotypage. Cependant, des biais technologiques - qualité du génotypage, efficacité des sondes variable selon l'allèle, etc... - peuvent alors influencer le recueil des données.

Un biais de confusion conduit à la détection d'une association statistiquement significative, mais qui en réalité est le fruit de l'influence d'un facteur externe appelé facteur confondant, associé respectivement avec le facteur d'exposition et la maladie étudiée, mais n'est pas une conséquence de l'exposition. D'autre part, la prise en compte du facteur confondant modifie l'estimation de l'effet du facteur d'exposition. Par exemple, on considère une étude d'association qui conclut que la consommation d'alcool augmente le risque de cancer du poumon. Or, les individus qui fument consomment généralement de l'alcool, et le tabagisme n'est pas une conséquence de la consommation d'alcool. Dans ce contexte, il semble pertinent de considérer le tabagisme comme un potentiel facteur confondant et de le prendre en compte dans le design de l'étude.

II.4 .2 Analyse statistique

La table 2 présente les notations utilisées pour désigner les effectifs des individus inclus dans l'étude en fonction de leur statut - cas M^+ ou témoin M^- - et de leur exposition - exposé E^+ ou non-exposé E^-.

	Cas M^+	**Témoins M^-**
Exposés E^+	a	b
Non-exposés E^-	c	d
Total	$a+c$	$b+d$

Table 2 – Table de contingence contenant les notations des effectifs des individus inclus dans une étude de type cas/témoins en fonction de leur statut et de leur exposition

L'indicateur statistique utilisé dans une étude cas/témoin est l'odds ratio, noté OR, appelé également rapport des cotes. Cet indicateur est le rapport des odds de développer la maladie lorsque l'on est respectivement exposé ou non. L'odds - ou la cote - d'un évènement de probabilité p est le rapport de la probabilité que cet évènement se produise sur la probabilité qu'il ne se produise pas, soit $Odds = \frac{p}{1-p}$. Prenons par exemple le cas du lancement d'une pièce, et étudions l'odds de l'évènement $E =$ « La pièce tombe côté face ». Si la pièce est parfaitement équilibrée, $p = 0.5$, et $Odds_E = \frac{0.5}{1-0.5} = 1$. Il y a autant de chances que la pièce tombe sur « pile » que sur « face ».

L'odds ratio de la maladie se calcule d'après la formule suivante :

$$OR = \frac{Odds_M^{E+}}{Odds_M^{E-}} = \frac{\frac{P(M^+|E^+)}{P(M^-|E^+)}}{\frac{P(M^+|E^-)}{P(M^-|E^-)}}$$

avec :

- $Odds_M^{E+}$: Rapport des cotes de développer la maladie chez les exposés.
- $Odds_M^{E-}$: Rapport des cotes de développer la maladie chez les non exposés.

Or, les risques absolus de maladie en fonction de l'exposition, soit les probabilités $P(M^+|E^+)$, $P(M^-|E^+)$, $P(M^+|E^-)$ et $P(M^-|E^-)$, sont des valeurs biaisées par le choix arbitraire du nombre de cas et de témoins dans l'étude, et ne doivent pas être estimées dans une étude cas/témoins. On peut cependant utiliser le théorème de Bayes pour exprimer différemment le calcul de l'OR, en fonction de probabilités d'exposition en fonction du statut cas/témoin qui ne sont cette fois pas biaisées par les effectifs arbitraires des cas et témoins.

Formule de Bayes : $$P(A|B) = \frac{P(A).P(B|A)}{p(B)}$$

Ce qui conduit à :

$$OR = \frac{Odds_M^{E+}}{Odds_M^{E-}} = \frac{\frac{P(M^+|E^+)}{P(M^-|E^+)}}{\frac{P(M^+|E^-)}{P(M^-|E^-)}} = \frac{\frac{P(E^+|M^+)}{P(E^-|M^+)}}{\frac{P(E^+|M^-)}{P(E^-|M^-)}} = \frac{Odds_E^{M+}}{Odds_E^{M-}} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{a.d}{b.c}$$

avec :

- $Odds_E^{M+}$: Rapport des cotes d'être exposé chez les cas
- $Odds_E^{M-}$: Rapport des cotes d'être exposé chez les témoins.

L'odds ratio associé à un facteur donné s'interprète comme un facteur multiplicateur du risque qu'un individu exposé à ce facteur a de présenter l'évènement étudié par rapport à un individu non exposé. On ne peut pas conclure sur la significativité de l'OR seul, il ne peut s'interpréter qu'avec son intervalle de confiance à 95%, noté IC95%. La valeur obtenue pour

l'OR n'est qu'une estimation, et l'IC95% est l'intervalle dans lequel on est sûr à 95% que la valeur réelle de l'OR se situe. Le calcul de l'intervalle de confiance de l'odds ratio se fait par l'intermédiaire de l'intervalle de confiance de son logarithme népérien, car celui-ci est distribué selon une loi normale et il est possible d'estimer sa variance :

$$Var(ln(OR)) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

Ce qui donne pour les bornes inférieures et supérieures de l'intervalle de confiance :

$$IC95\%_{Inf} = e^{\,ln(OR) - 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}} \quad \text{et} \quad IC95\%_{Sup} = e^{\,ln(OR) + 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}$$

On conclut sur la significativité de l'association étudiée en fonction des valeurs de l'odds ratio et de son intervalle de confiance.

Si 1 $\in$ *IC*95% : on ne peut statistiquement pas dire que $OR \neq 1$, on conclut qu'il n'y a statistiquement pas d'association entre le risque de maladie et le facteur d'exposition.

Si 1 $\notin$ *IC*95% et $OR > 1$: il existe une association statistiquement significative. $OR > 1$ signifie que le facteur d'exposition est associé avec une augmentation du risque de voir se réaliser l'évènement, dans notre cas de développer la maladie étudiée. Un individu exposé voit son risque de développer la maladie multiplié par OR.

Si 1 $\notin$ *IC*95% et $OR < 1$: il existe une association inverse statistiquement significative. $OR < 1$ signifie que le facteur d'exposition diminue le risque de voir se réaliser l'évènement. Un individu exposé voir son risque de développer la maladie multiplié par OR, soit - puisque $OR < 1$ - divisé par $\frac{1}{OR}$.

Cette méthode de calcul de l'odds ratio est utilisé en cas d'analyse univariée, où l'influence d'un seul facteur, le facteur d'exposition est pris en compte. Cependant, il est fréquent de vouloir estimer un odds ratio tout en tenant compte d'autres facteurs tels que l'âge, le sexe, le pays d'origine, etc... Cela permet d'ajuster l'effet du facteur d'exposition en s'affranchissant du potentiel effet qu'ont les autres facteurs sur la probabilité de l'évènement étudié. Ce type d'analyse est appelé multivarié. L'estimation de l'odds ratio est alors plus complexe, et nécessite d'utiliser un modèle appelé régression logistique.

Un modèle logistique représente, à l'aide d'une fonction de lien appelée *logit*, la probabilité p d'un évènement comme une combinaison linéaire d'une ou plusieurs covariables dont on cherche à estimer les coefficients :

$$logit(p) = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

avec :

- β_0 : intercept du modèle.
- x_1 .. x_n : covariables dont on veut tester ou ajuster l'effet sur le risque.
- β_1 .. β_n : coefficients des covariables inclues dans le modèle, que l'on cherche à estimer.

On utilise cette fonction de lien car la variable de réponse p ainsi transformée devient linéaire sur l'échelle du logarithme népérien. L'interprétation des coefficients, estimés sur cette échelle nécessite donc d'en prendre l'exponentielle. e^{β_0} s'interprète comme l'odds de la maladie en absence d'exposition ou dans la catégorie de référence. e^{β_i} est l'odds ratio associé à la covariable x_i et son interprétation numérique dépend du codage de la variable x_i (numérique, catégoriel, binaire, etc...).

Prenons pour exemple une analyse dans laquelle on souhaite modéliser le risque p de cancer du poumon en fonction de l'âge et du fait d'être ou non fumeur, et d'avoir ou non du diabète. On dispose des données suivantes : l'âge de l'individu en années, son statut tabagique (NF : non-fumeur, AF : ancien fumeur, ou F : fumeur), et l'éventuelle existence de diabète chez cet individu. On construira donc le modèle suivant :

$$logit(p) = \beta_0 + \beta_a x_a + \beta_{AF} x_{AF} + \beta_F x_F + \beta_d x_d$$

avec :

x_a : Covariable représentant l'âge de l'individu, variable numérique discrète, en années.

x_{AF}, x_F : Covariables correspondant au statut tabagique de l'individu. Indicatrices de classe valant 1 si l'individu fait partie de la catégorie correspondante, 0 sinon. On n'introduit pas de covariable pour le statut non fumeur (NF), car cette catégorie est choisie comme catégorie de référence pour le statut tabagique.

x_d : Covariable indiquant si l'individu a du diabète ou non. Indicatrice de classe, codée 1 en présence de diabète, 0 sinon.

β_i : Paramètres du modèle à estimer.

Pour un modèle linéaire classique, on estime les paramètres du modèle par la méthode des moindres carrés. Or, dans un modèle logistique, les résidus ne suivent pas une loi normale. Pour estimer les paramètres du modèle, on utilise alors la méthode du maximum de vraisemblance. La vraisemblance d'un modèle est la probabilité que les estimations effectuées des paramètres de ce modèle reflètent la réalité sachant les données que l'on a observé. Prenons un exemple simple. On lance une pièce 1000 fois. Sur ces 1000 lancers, on obtient 745 fois pile, et 255 fois face. Le nombre de lancers où la pièce est retombée côté pile se modélise par une loi binomiale $\mathcal{B}(n, p)$ où $n = 1000$ puisque l'on réalise n lancers, et où on cherche à estimer $p = p_{pile}$. À la vue de ces résultats, il est peu probable que la pièce soit équilibrée, et que $p_{pile} = p_{face} = \frac{1}{2}$. Il est cependant beaucoup plus **vraisemblable** à la vue des données que $p_{pile} = 0.745$ et que

$p_{face} = 0.255$. Cette manière d'estimer les paramètres du modèle est appelée méthode de maximum de vraisemblance, et consiste à prendre pour estimation des paramètres la valeur des paramètres la plus probable à la vue des données observées. L'estimation des paramètres par maximisation de la vraisemblance peut être effectuée de manière très rapide à l'aide de logiciels de statistiques tels que R, à l'aide de la fonction `glm`.

Pour notre exemple d'étude sur le cancer du poumon, on suppose avoir effectué l'estimation des paramètres des covariables ainsi que le calcul de leur intervalle de confiance, et obtenu les résultats suivants :

Paramètre β_i	**Estimation de β_i**	**Estimation de $OR_i = e^{\beta_i}$**	**IC95%**
β_a	0.006	1.006	1.002 - 1.011
β_{AF}	0.42	1.522	1.136 - 2.054
β_F	1.535	4.641	3.265 - 5.945
β_d	-0.038	0.962	0.783 - 1.013

Interprétation de β_a : 1 est exclu de l'intervalle de confiance β_a. L'estimation effectuée de e^{β_a} est donc statistiquement différente de 1 : le risque de cancer du poumon est multiplié par 1.006 par année d'âge supplémentaire. Ainsi, à niveau égal des autres facteurs considérés, un individu âgé de 70 ans a un risque de cancer du poumon multiplié par $e^{10\times\beta_a}$ par rapport à un individu âgé de 60 ans.

Interprétation de $e^{\beta_{AF}}$ et e^{β_F} : De la même manière, les OR $e^{\beta_{AF}}$ et e^{β_F} sont statistiquement différent de 1. La classe de référence pour cette variable étant les non fumeurs NF, un individu fumeur a 4.641 fois plus de chances de cancer du poumon qu'un non fumeur, et un ancien fumeur a 1.522 fois plus de chances de cancer du poumon qu'un non fumeur, à niveau égal des autres covariables.

Interprétation de β_d : L'intervalle de confiance de e^{β_d} n'exclut pas 1, on ne peut pas dire de manière statistiquement significative que le diabète a une influence sur le risque de cancer du poumon. Il est inutile de conserver ce facteur dans l'expression du modèle.

On peut également comparer le risque de cancer du poumon entre deux personnes en fonction des différentes covariables. Par exemple, calculons l'OR d'une personne P_1 âgée de 62 ans et ancienne fumeuse par rapport à une personne P_2 de 55 ans, fumeuse :

- Pour P_1 : $\quad logit(p)_{P_1} = \beta_0 + 62.\beta_a + \beta_{AF}$
- Pour P_2 : $\quad logit(p)_{P_2} = \beta_0 + 55.\beta_a + \beta_F$

On calcule l'odds ratio de P1 par rapport à P2 :

$$OR_{P1/P2} = \frac{e^{logit(p)_{P_1}}}{e^{logit(p)_{P_2}}} = \frac{e^{\beta_0}.e^{62.\beta_a}.e^{\beta_{AF}}}{e^{\beta_0}.e^{55.\beta_a}.e^{\beta_F}} = e^{7.\beta_a+\beta_{AF}-\beta_F} = 0.342$$

Ainsi, d'après les estimations du modèle, un ancien fumeur de 62 ans a un risque de cancer du poumon multiplié par 0.342, soit divisé par $\frac{1}{0.342} = 2.92$ par rapport à une personne fumeuse

âgée de 55 ans.

Il est possible d'attester de la significativité d'un, de plusieurs, ou de l'ensemble des coefficients des facteurs inclus dans un modèle de manière plus formelle en utilisant le test de Wald, le test du score, ou le test du rapport de vraisemblance. Ces trois tests sont équivalents asymptotiquement. Ils testent les deux hypothèses suivantes :

H_0 : $\beta = \beta_0 \quad (= 0) \qquad$ avec $\beta = 1 \,..\, K$

H_1 : Au moins 1 $\beta \neq 0$

Dans la plus grande partie des cas, on teste si les coefficients β sont nuls ou non, mais il est possible de tester leur égalité à d'autres valeurs de référence.

Expression des statistiques des tests cités pour un modèle multiparamétrique :

– Test de Wald :

$$(\widehat{\beta} - \beta_0)^T . \, I(\widehat{\beta})^{-1} \, (\widehat{\beta} - \beta_0) \qquad \hookrightarrow X^2_K$$

– Test du Rapport de Vraisemblance :

$$2 \left(LV(\widehat{\beta}) - LV(\beta_0) \right) \qquad \hookrightarrow X^2_K$$

– Test du score :

$$U^T.(\beta_0).I(\beta_0)^{-1}.U(\beta_0) \qquad \hookrightarrow X^2_K$$

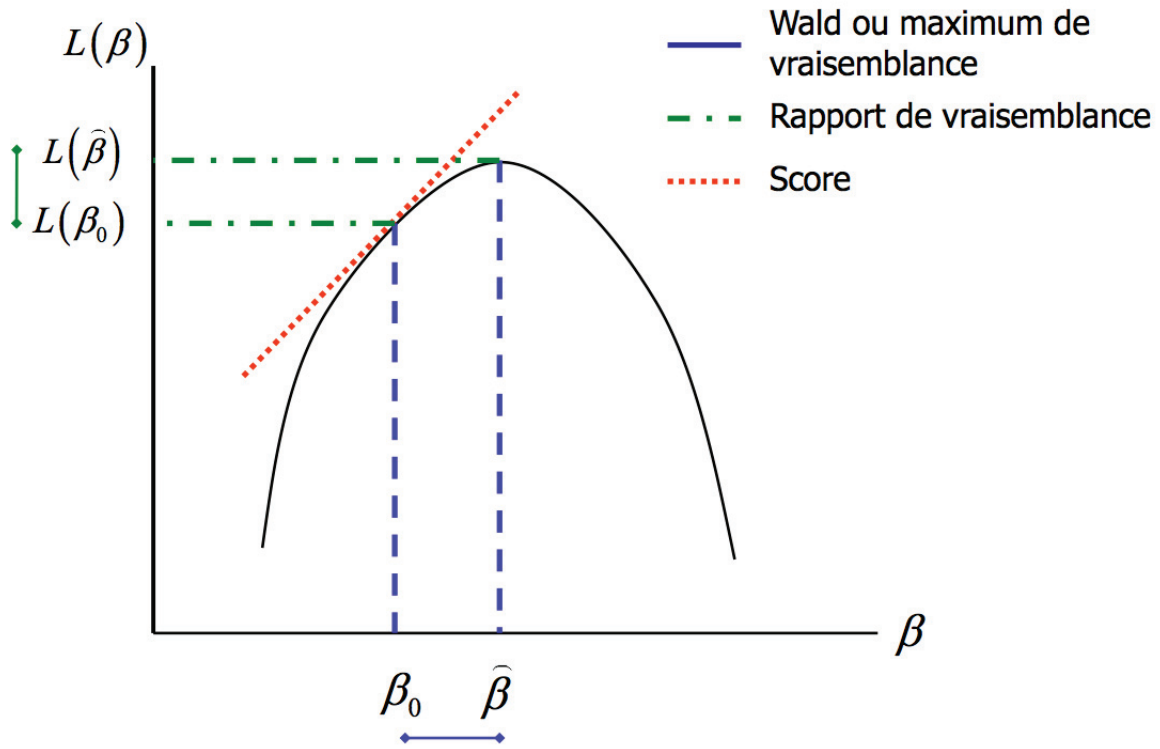
où I est la matrice d'information de Fisher :

$$I(\beta_0) = -E \left(\frac{\partial^2 \, ln \, L(\beta_0)}{\partial \beta \, \partial \beta'} \right)$$

La figure 14 représente graphiquement la différence de principe entre les trois tests cités, tous basés sur l'estimation de la vraisemblance. Le test du score revient à tester que la pente en β_0 est nulle, le test de Wald que l'écart entre $\widehat{\beta}$ et β_0 en abscisse est nul, et le test du rapport de vraisemblance que l'écart entre la log-vraisemblance en $\widehat{\beta}$ et en β_0 en ordonnées est nul.

La réalisation de chacun de ces tests permet d'obtenir une p-value qui permet de conclure statistiquement en fonction du risque de première espèce α toléré. Dans une étude d'association génétique, il est fréquent que plusieurs dizaines de marqueurs génétiques soit testés. Généralement, un modèle par marqueur est construit, avec $\alpha = 5\%$. Or, lorsque plusieurs modèles sont estimés, le risque d'obtenir des faux-positifs augmente, phénomène appelé inflation de α. Si on multiplie le tirage d'une variable aléatoire k fois avec à chaque fois un risque de première espèce $alpha = 5\%$, alors on obtient un risque de faux-positif global $\alpha_k = 1 - (1 - \alpha)^k$. Pour $k = 80$, $\alpha_{80} > 0.98$. Ainsi, en ayant réalisé 80 tests, on a 98% de chances d'obtenir au moins 1 faux positif. Afin de ne pas conclure à tort à l'effet d'un facteur sans pour autant limiter le design des études mises en place, il convient d'utiliser des méthodes statistiques afin de contrôler le risque α global dans le cadre de tests multiples. Les méthodes de correction les plus connues sont la

Figure 14 – Principe des trois tests usuels permettant de conclure sur la significativité de $\widehat{\beta}$



Représentation graphique de la vraisemblance $L(\beta)$ en fonction de β. Le principe des tests de Wald, du rapport de vraisemblance et du score sont représentés schématiquement.

méthode de correction de Bonferroni, et la méthode de correction de Benjamini-Hochberg.

La méthode de correction de Bonferroni est basée sur le contrôle du *Family Wide Error Rate* - ou FWER - et contrôle le risque de n'avoir aucun faux-positif. Le principe de cette correction est d'être plus stringent sur chaque test unitaire réalisé afin que le risque global $\tilde{\alpha}$ reste inférieur à 5%. Pour cela, on divise le risque unitaire α par le nombre de tests réalisés, pour conserver $\tilde{\alpha} = \frac{\alpha}{n} < 5\%$. La correction de Bonferroni est très conservative. Au final, elle n'est capable de détecter que très peu de vrais positifs. En terme de p-value, cette correction s'applique également en calculant les p-values ajustées $\tilde{p}$, c'est à dire en multipliant les p-values p obtenues par les tests unitaires par le nombre total de tests réalisés : $\tilde{p} = np$. Elles s'interprètent alors de la manière habituelle, en comparaison du seuil $\tilde{\alpha}$ fixé le plus souvent à 5%.

La méthode de correction de Benjamni-Hochberg[104] est basée sur le contrôle du *False discovery Rate* - ou FDR, et contrôle la proportion de faux-positifs. La correction de Benjamini-Hochberg a tendance à conserver plus de vrais positifs, et à être moins stringente. Si FD est le nombre de faux-positifs détectés par l'expérience, et n le nombre total de tests réalisés, alors elle garantit : $\frac{FD}{n} < \alpha$.

II.4 .3 Méta-analyses

Il est fréquent de grouper plusieurs études cas-témoins effectuées sur des populations indépendantes afin d'augmenter les effectifs pour maximiser la puissance statistique des tests d'association réalisés, ou de conclure sur l'effet d'un facteur lorsque plusieurs études sont discordantes ; c'est ce qu'on appelle une méta-analyse. Une méta-analyse consiste à regrouper les données et/ou les résultats de ces différentes études afin d'estimer au plus juste l'odds ratio de l'association testée. Cependant, l'effet des covariables peut être hétérogène entre les études inclues dans la méta-analyse.

Il existe plusieurs manières de quantifier l'hétérogénéité dans une méta-analyse. Les statistiques Q et I^2 permettent respectivement de détecter la présence ou l'absence d'hétérogénéité entre plusieurs études, et de quantifier celle-ci :

$$Q = \sum_{i=1}^{K} \frac{(\beta_i - \widehat{\beta})^2}{\sigma_i^2} \quad \hookrightarrow X^2_{K-1} \qquad \text{et} \qquad I^2 = \frac{Q - df}{Q} \times 100$$

avec :

- K : Nombre d'études inclues dans la méta-analyse.
- $\widehat{\beta}$: estimation de l'effet conjoint par maximum de vraisemblance basée sur toutes les études (modèle à effets fixes).
- β_i : estimation de l'effet β dans l'étude i.
- σ_i^2 : variance au sein de l'étude i.

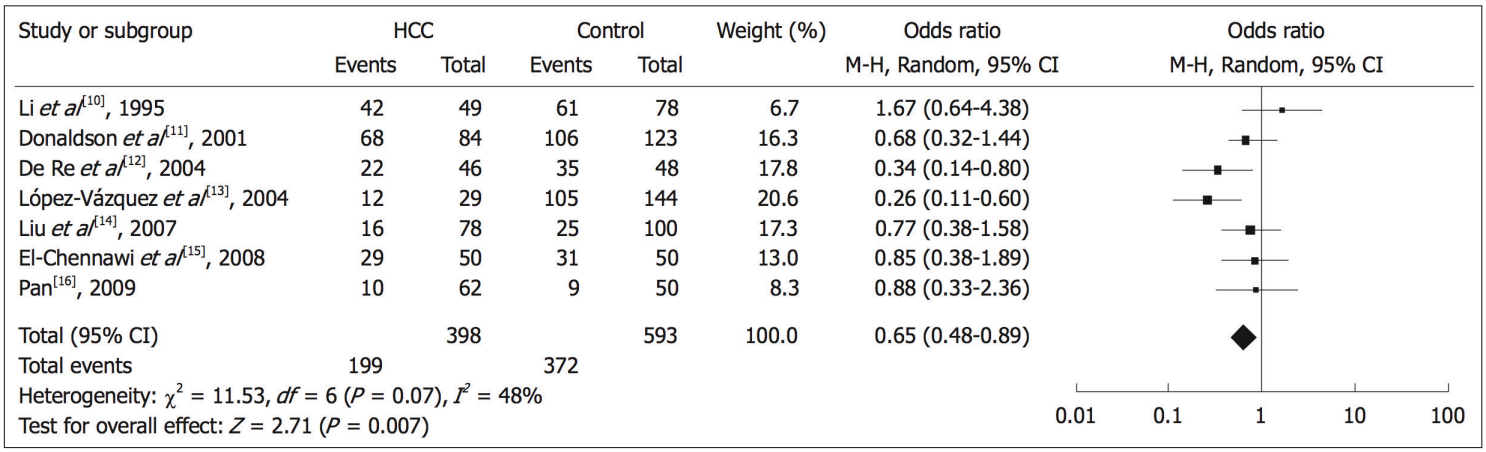
Q permet d'établir si la variabilité observée dans chaque étude est plus importante qu'attendu par hasard. Si Q s'éloigne trop d'une distribution du X^2 à $K-1$ degrés de liberté, il existe une hétérogénéité statiquement significative entre les effets estimés dans chaque étude. Cependant, Q a une puissance statistique faible pour détecter l'hétérogénéité lorsque K est faible, comme c'est souvent le cas dans les méta-analyses. A contrario, Q est souvent trop sensible lorsque K est élevé. De plus, Q permet de conclure sur l'existence d'une éventuelle hétérogénéité, mais pas de la quantifier si elle existe. Pour cela, on utilise I^2. I^2 représente le pourcentage de la variabilité observée qui est due à l'existence d'hétérogénéité plutôt qu'au hasard. Si 0 est inclu dans l'intervalle de confiance de I^2, alors on peut conserver l'hypothèse d'homogénéité entre les études[105]. Dans le cas contraire, on peut qualitativement qualifier le taux d'hétérogénéité de faible pour $I^2 < 25$ %, de moyen pour $25\ \% < I^2 < 75$ %, et de fort pour $I^2 > 75$ %[105].

Lorsque de l'effet estimé entre les études est hétérogène, il est nécessaire de construire un modèle à effets mixtes à la place d'un modèle à effets fixes. Dans un modèle à effets mixtes, certains effets sont modélisés selon une variable aléatoire suivant une loi normale de moyenne nulle et de variance à estimer. Autrement dit, on prend en compte dans la modélisation plus de variabilité dans l'effet observé de certains facteurs pour certaines études. L'estimation obtenue

tiendra compte de cette variabilité, l'intervalle de confiance de l'effet combiné sera plus large qu'avec le modèle fixe.

Il est fréquent de représenter les résultats d'une méta-analyse à l'aide d'un graphe appelé *forest-plot* ou « graphique en forêt ». La figure 15 représente le forest-plot d'une méta-analyse testant l'association entre le risque de carcinome hépatocellulaire et l'allèle DQB1*03[106], polymorphisme localisé sur la séquence codante d'un antigène leucocytaire humain - ou HLA pour *Human Leukocyte Antigen*. 7 études ont été sélectionnées et inclues dans cette méta-analyse. L'OR de chaque étude est représenté par un carré noir dont la taille est proportionnelle aux effectifs de l'étude. L'intervalle de confiance de chaque OR est représenté par un segment noir horizontal. L'OR global estimé est représenté par un losange noir, et s'interprète avec son intervalle de confiance comme un OR classique. 5 des 7 études inclues n'ont pas permis d'estimer un OR statistiquement significatif, puisque leur intervalle de confiance contient la valeur 1. La proportion de variabilité due à une hétérogénéité entre les études a été estimée à $I^2 = 48\%$. Un modèle incluant des effets aléatoires a donc été construit, et a conduit à une estimation de l'OR global de 0.65, pour un intervalle de confiance à 95% de 0.48-0.89. Ainsi, d'après les résultats de cette méta-analyse, les individus porteurs de l'allèle HLA-DQB1*03 ont un risque de carcinome hépatocellulaire multiplié par 0.65, soit divisé par 1.54 environ, par rapport aux porteurs d'un autre allèle.

Figure 15 – Exemple de forest-plot représentant graphiquement les résultats d'une méta-analyse



Cette méta-analyse étudie l'association entre le risque de carcinome hépatocellulaire et l'allèle DQB1*03, polymorphisme de la séquence codante d'un antigène de leucocytes humain - ou HLA pour *Human Leukocyte Antigen*. L'OR combiné est significatif 0.65 (IC95% 0.48-0.89), l'allèle HLA-DQB1*03 est statistiquement inversement associé avec le risque de carcinome hépatocellulaire. *D'après Xin et. al., World Journal of Gastroenterology, 2011* [106]

II.4 .4 Analyse gène candidat

Les analyses dites gène candidat sont un type d'étude dont le principe est de rechercher une association entre la pathologie et les polymorphismes situés dans et à proximité de gènes qui pourraient être impliqués de manière fonctionnelle dans le développement de la maladie. Cette approche suppose d'avoir des connaissances ou tout du moins des hypothèses solides sur d'une part, les altérations fonctionnelles conduisant à l'apparition du phénotype pathologique, et d'autre part sur les gènes potentiellement impliqués, leur localisation, leur structure, leur fonction, et leurs polymorphismes.

Le choix des gènes candidats est effectué d'après les connaissances déjà disponibles pour l'investigateur de l'étude. Une analyse poussée de la bibliographie peut permettre d'émettre des hypothèses quant à l'héritabilité du phénotype pathologique, la pénétrance de la maladie. Des études de liaison génétique peuvent pousser l'investigateur à s'intéresser à certaines régions génomiques en particulier. L'étude de l'expression génique peut conduire à sélectionner des gènes différentiellement exprimés entre une condition contrôle et une condition pathologique. de plus, il est utile de prendre en compte les résultats obtenus sur les gènes homologues d'espèces proches et qui peuvent potentiellement être extrapolés à l'Homme. Enfin, on sait que dans certaines pathologies des mécanismes de régulation complets sont altérés, que ce soit au niveau des voies de signalisation, des réseaux de régulation métaboliques ou d'expression. Les gènes impliqués dans ces réseaux sont également de bons candidats fonctionnels.

Bien qu'un nombre limité de gènes soient sélectionné pour une analyse gène candidat, tous les SNPs situés à proximité de ces gènes ne sont généralement pas génotypés. Il existe peu de régions génomiques qui soient extrêmement conservées et dans lesquelles peu de polymorphismes ont été identifiés. Une étape de classement hiérarchique par priorité des SNPs connus est souvent effectuée pour aider l'investigateur dans ce choix. Un des premiers critères utilisés est la fréquence allélique des SNPs. En effet, les SNPs que l'on cherche à détecter par des études d'association ont généralement un effet modéré à faible. Or, si ces SNPs ont une fréquence allélique très faible, alors de manière générale, les études mises en place n'incluent pas assez d'individus pour que l'on soit capable de détecter de manière statistiquement significative l'association si elle existe. Il est donc inutile de les génotyper. Le seuil de fréquence allélique utilisé dans la sélection des SNPs varie entre 5% et 1%, en dessous duquel les SNPs sont considérés comme rares. A contrario, les polymorphismes très fréquents dans la population générale, dont la fréquence de l'allèle mineur - ou MAF pour *Minor Allele Frequency* - est proche de 0.5 sont généralement peu informatifs.

Il également important de sélectionner des SNPs qui soient adaptés à la population dans laquelle l'étude sera réalisée, en particulier en terme d'appartenance ethnique. En effet, certains polymorphismes n'ont été observés que dans certaines populations, et sont extrêmement rares dans la population dans laquelle les individus génotypés seront échantillonnés, il est inutile de les génotyper. Par exemple, d'après le projet des 1000 génomes[107] dont l'objectif est de caractériser les polymorphismes du génome humain, et ce dans différents groupes ethniques et populations, le SNP `rs8176312`, situé sur le gène *BRCA1*, est observé avec une MAF de 0.38 au sein de la population africaine. Cependant, les populations européennes, est-asiatiques, et américaine d'origine caucasienne, la MAF est respectivement de 0.008, 0, et 0.03. Ainsi, alors

que ce SNP est présent chez environ 40% de la population africaine, il est observé chez environ 3% des Américains, chez moins de 1% des Européens, et n'est pas présent au sein de la population d'origine est-asiatique.

Certains SNPs ont été observés dans l'ensemble des populations étudiées, mais la fréquence de leurs allèles peut être très différente entre ces populations. Ainsi, le snp `rs43182741`, également situé sur le gène *BRCA1*, a été étudié chez les Africains, chez les Européens, chez les Américains d'origine caucasienne, et chez les populations d'Asie de l'est. Ce SNP est biallélique, et les allèles observées sont A et T. L'allèle mineur observé dans toutes ces populations est l'allèle A, à l'exception de la population d'origine africaine, pour laquelle l'allèle mineur est T, avec une MAF de 0.128. La MAF de ce SNPs dans les trois autres populations évoquées précédemment est respectivement 0.367, 0.456, 0.330. Ainsi, en moyenne l'allèle A est porté par 87% des Africains, 46% des Américains, 37% des Européens, et 33% des Asiatiques de l'est.

Enfin, un des critères les plus importants pour sélectionner les SNPs qui seront génotypés est leur potentiel impact biologique. Comme évoqué précédemment, un SNP situé dans une région codante peut être synonyme lorsqu'il n'engendre pas de modification de l'acide aminé qu'il encode, faux-sens lorsqu'il engendre un changement d'acide aminé, ou encore non-sens lorsqu'il engendre une terminaison prématurée de l'ARN messager correspondant et conduit à la synthèse d'une protéine tronquée. Ces deux dernières catégories de SNPs font partie de celles ayant un impact fonctionnel le plus probable. Les SNPs synonymes ou ceux situés hors des régions codantes ont moins de chances d'avoir un impact fonctionnel et d'être associé à la maladie. En effet, 80% à 90% des mutations pathogènes situées sur *BRCA1* et à l'origine d'un cancer du sein sont des mutations tronquantes conduisant à la synthèse d'une protéine incomplète ou inexistante[108]. Cependant, des SNPs situés dans des régions régulatrices peuvent avoir un impact tout aussi fort, en modifiant la régulation de l'expression du gène, ou encore l'épissage, la stabilité ou la localisation de l'ARN messager. De plus, un SNP situé dans une région génomique ultraconservée entre plusieurs espèces a de plus forte chances d'avoir un impact fonctionnel sous l'hypothèse que cette région est soumise à une pression de sélection positive, et que son rôle confère un avantage évolutif. Il existe des algorithmes tels que SIFT[109] ou PolyPhen[110] qui prennent en compte l'ensemble de ces paramètres afin d'effectuer des prédictions sur l'impact fonctionnel qu'un polymorphisme peut avoir. SIFT qualifie les polymorphismes de bénins, potentiellement néfaste, ou probablement néfaste. PolyPhen les qualifie de tolérés ou délétères. Les deux algorithmes calculent un score qui peut aider à classer ces polymorphismes. Enfin, il est souvent utile d'avoir au moins un à plusieurs SNPs au sein d'un bloc de déséquilibre de liaison selon sa taille lorsque la structure du LD est connue pour les gènes candidats.

Le principal défaut des études de type gène candidat est le besoin d'avoir déjà des éléments laissant à penser qu'un gène donné pourrait être impliqué de manière fonctionnelle dans le développement de la maladie. Or il existe de très nombreuses maladies multifactorielles dont nous commençons tout juste à comprendre les mécanismes de développement les plus simples.

Parmi les gènes de prédisposition au cancer du sein identifiés par une approche gène candidat, on trouve *TP53*, *PALB2*, *CHEK2*, *ATM*, et *BRIP1*. D'autres polymorphismes, ayant un effet moindre ont également été identifiés comme facteurs de susceptibilité au cancer du

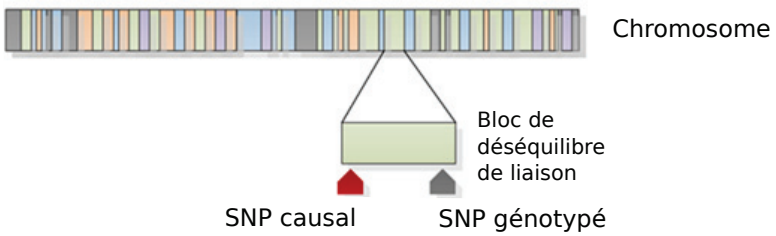
sein via des approches gène candidat[111]. Parmi l'ensemble des mutations et polymorphismes identifiés via ce type d'approche, certains sont des variants relativement fréquents dans la population, alors que d'autres sont beaucoup plus rare. On distingue généralement les SNPs de prédisposition, qui ont une fréquence relativement élevée, des variants rares. Les plus récentes études gène candidat[112], bien qu'incluant un très grand nombre d'individus, ont commencé à obtenir des résultats d'association négatifs, suggérant que cette approche avait atteint sa limite, et que d'autres types de méthodes étaient requises afin de détecter des polymorphismes dont la pénétrance et l'effet étaient plus faibles.

II.4 .5 Études pangénomiques, ou *Genome Wide Association Studies*

Les études pangénomiques - ou GWAS pour *Genome Wide Association Study* - sont une méthode de détection d'association publiée pour la première fois en 2005. La première étude[113] de ce type réalisée portait sur la susceptibilité génétique à la dégénérescence maculaire liée à l'âge, ou DMLA. Cette approche est une généralisation de l'approche gène candidat. Alors que les études du type gène candidat ciblent des gènes dont la fonction était connue et qui pouvaient potentiellement être impliqués de manière causale et fonctionnelle dans le développement de la maladie étudiée, les études pangénomiques cherchent à détecter des associations à l'échelle du génome entier. Ainsi, des polymorphismes répartis sur l'intégralité du génome sont génotypés, et ce sur un grand nombre de cas et de témoins, afin de détecter quels polymorphismes sont associés ou inversement associés au phénotype pathologique.

Cependant, bien que l'on analyse les polymorphismes à l'échelle du génome entier, tous les polymorphismes connus du génome ne sont pas génotypés pour autant. L'étude de la structure des blocs de déséquilibre de liaison sur le génome (voir Section II.1) permet d'optimiser le choix des SNPs à génotyper. Du fait de l'existence de ces blocs qui traduisent le fait que certaines portions chromosomiques sont généralement transmises à la descendance sans subir de remaniement dû à la recombinaison, on estime qu'à partir du génotype d'un SNP, on peut inférer le génotype des SNPs qui sont en fort déséquilibre de liaison avec lui, c'est à dire $r^2 > 0.8$. Ainsi, comme illustré sur la figure 16, pour détecter un potentiel SNP causal, il suffit de génotyper au moins un SNP en haut déséquilibre de liaison avec lui.

Figure 16 – Détection indirecte d'une association



Au sein d'un même bloc de déséquilibre de liaison, le génotype du SNP causal peut être inféré à partir de celui d'un autre SNP, qui lui est génotypé, et en haut déséquilibre de liaison avec le SNP causal.

Dans une étude GWAS, le génotypage s'effectue généralement sur des puces commerciales dont les deux principaux fabricants sur le marché sont Affymetrix et Illumina. Toutes les puces GWAS n'ont pas la même densité, c'est à dire qu'elles ne permettent pas de génotyper le même nombre de SNPs. Parmi les puces les plus récentes, certaines permettent de génotyper beaucoup plus de SNPs que les premières générations de puces GWAS. Alors que la première étude pangénomique réalisée incluait environ 100 000 SNPs, les puces actuelles ont une densité pouvant aller jusqu'à 4 millions de SNPs. Contrairement aux puces de densité moyenne, soit environ 700 000 SNPs, les puces les plus denses permettent également de génotyper des SNPs dits rares, dont la fréquence de l'allèle mineur peut être inférieure à 1%. Au fil des années, notre connaissance de la structure du génome humain s'est également affinée, et les puces actuelles ciblent beaucoup mieux la variabilité de notre génome que les premières puces, même à nombre égal de marqueurs génotypés. Les puces les plus récentes sont de plus spécifiquement adaptées aux diverses populations ciblées (caucasienne, afro-américaine, asiatique, etc...).

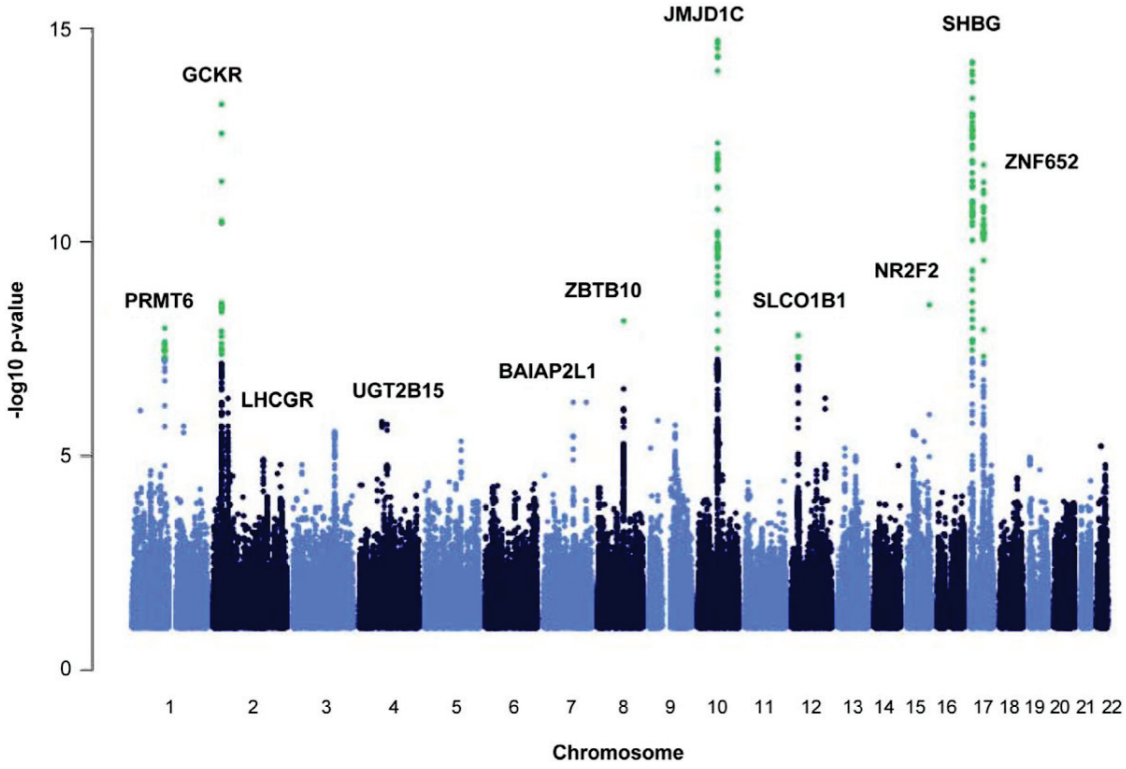
L'analyse statistique d'une étude pangénomique est très similaire à celle d'un étude gène candidat. Cependant, les niveaux de significativité espérés sont beaucoup plus stringents que ceux d'une étude d'association classique. En effet, tester simultanément l'effet potentiel de centaines de milliers de SNPs expose au risque d'obtenir un nombre potentiellement très élevé d'associations significatives qui ne sont dues en réalité qu'au hasard (Effet des tests multiples, voir Section II.4 .2). De ce fait, les associations considérées comme intéressantes à l'issue de l'analyse d'une étude pangénomique sont celles ayant une p-value inférieure à 10^{-6} voire à 10^{-8}. Il est usuel de représenter les résultats d'une étude pangénomique de manière visuelle en construisant un graphe de Manhattan, ou *Manhattan plot*, dont un exemple est représenté dans la figure 17.

Un point représente la p-value de l'association évaluée pour chacun des SNPs inclus dans l'étude. En abscisse, les points sont positionnés par chromosome et par position chromosomique sur leur chromosome respectif. En ordonnée est représenté l'opposé du logarithme en base 10 de la valeur de la p-value. Les p-values les plus faibles, c'est à dire les plus significatives, sont donc représentées dans la partie supérieure du graphe. Les points représentés en vert ont une p-value inférieure à 10^{-7}.

Sur ce graphe, on distingue en plusieurs régions de nombreux points alignés verticalement, qui traduisent qu'autant de SNPs, situés tous très proches géographiquement ont des associations fortement significatives. C'est le cas par exemple, des régions dans lesquelles se trouvent les gènes *GCKR*, *JMJD1C*, et *SHBG*. L'observation de ce profil est cohérent avec l'existence de blocs de déséquilibre de liaison dans lesquels se trouveraient les SNPs causaux responsables des associations détectées. Les SNPs génotypés dans ces blocs, en fort déséquilibre de liaison avec les SNPs causaux, révèlent de manière indirecte l'association avec ceux-ci.

Afin de s'affranchir de la possibilité que les associations détectées dans une étude pangénomique soient dues au hasard, celle-ci doit impérativement être répliquée dans une population indépendante. C'est également vrai dans le cadre des études gène candidat, mais le nombre des SNPs testés étant moindre, le risque de détecter une association due au hasard l'est également. C'est ce qu'on appelle une étude pangénomique multi-étapes. La première étape consiste à gé-

Figure 17 – Exemple de graphe de Manhattan



Étude des loci associés avec le taux de SHBG (globuline se liant aux hormones sexuelles). Chaque point représente le résultat d'une association, c'est à dire la p-value obtenue pour chaque SNP testé. L'opposé du logarithme en base 10 de la p-value de chaque association est représenté en fonction de la position génomique sur leur chromosome respectif. Les points verts représentent les polymorphismes pour lesquels la p-value d'association est inférieure à 10^{-7}. *Coviello et al. Plos Genet. 2012*[114]

notyper un très grand nombre de SNPs sur un petit échantillon d'individus. On sélectionne alors les SNPs suffisamment fréquents dans la population d'intérêt. Du fait de l'augmentation permanente de nos connaissances sur la structure du génome et sur la fréquence des SNPs dans les populations humaines, cette étape est de moins en moins souvent mise en oeuvre. L'étape suivante consiste à génotyper les quelques centaines de milliers de SNPs retenus sur un grand échantillon de cas et de témoins. L'analyse de ces données fournit alors une mesure de l'association pour chaque SNP testé à cette étape, qu'il est possible de classer de la plus forte à la plus faible, c'est à dire en fonctions des p-values croissantes. Enfin, en fonction du budget alloué, on peut alors répliquer les associations les plus fortes en génotypant ces SNPs dans une population indépendante. Si l'association est également retrouvée lors de la réplication, alors celle-ci est estimée robuste et réelle.

Afin de réduire les coûts élevés de génotypage dans une étude pangénomique, une méthode consiste à imputer une partie des génotypes. Cette technique consiste à génotyper à l'aide de puces à forte densité une faible partie de la population des individus inclus dans l'étude, tandis que la majeure partie des individus sont génotypés sur une puce de moyenne densité. L'obser-

vation des combinaisons alléliques et des haplotypes dans la première partie de l'échantillon permet alors d'imputer de manière probabiliste pour la seconde partie de l'échantillon les génotypes des polymorphismes présents uniquement sur les puces à haute densité.

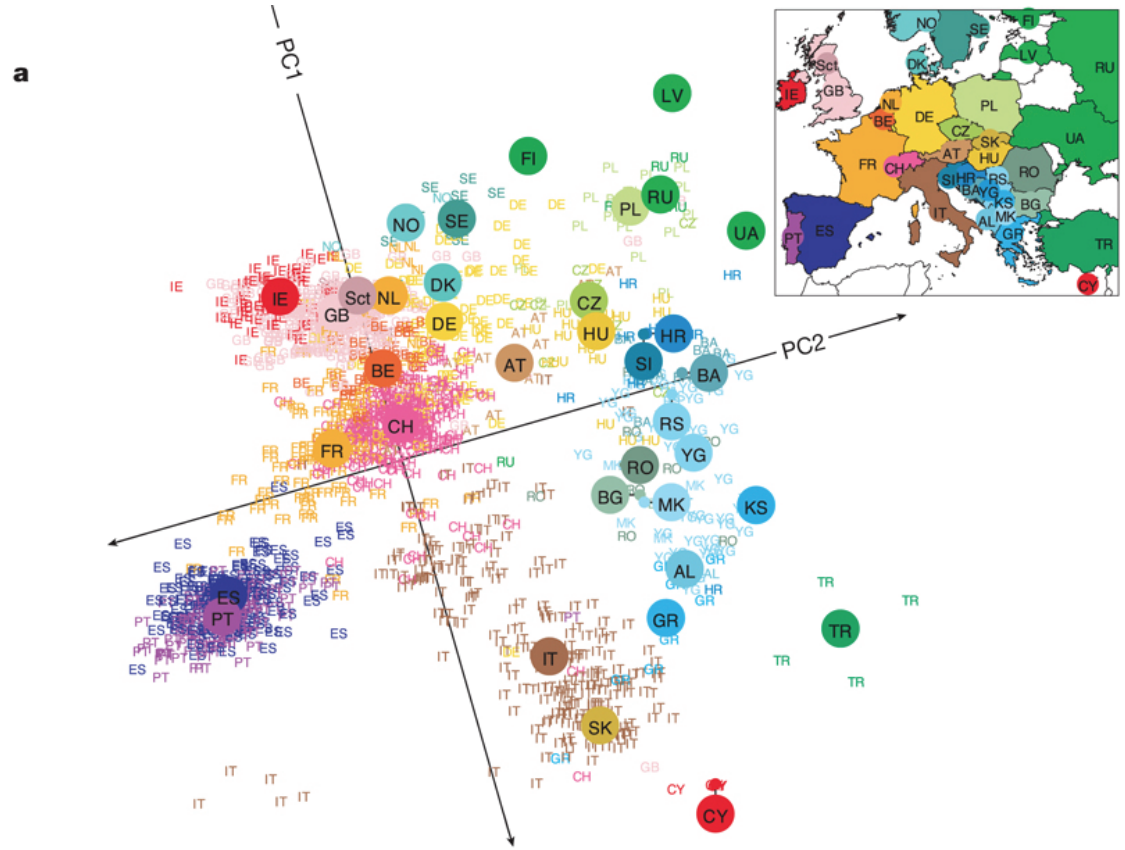
Le principal intérêt d'une étude pangénomique est l'absence d'*a priori* sur la fonction des polymorphismes testés. On peut ainsi identifier, notamment pour les maladies dont l'étiologie est mal connue, quels polymorphismes sont associés avec une modification du risque pour cette pathologie, et ainsi fournir des pistes de recherches fonctionnelles à explorer. On peut ainsi identifier des polymorphismes dans des gènes qui avaient *a priori* aucun lien fonctionnel avec la maladie, que l'on aurait donc jamais étudiés via une approche gène candidat. De plus, les études pangénomiques permettent d'identifier des effets faibles, avec des odds ratios associés inférieurs à 2 dans la majeure partie des cas, et souvent inférieurs à 1.5, qu'il était beaucoup plus difficile de détecter via des approches de liaison génétique.

Cependant cette approche, bien que révolutionnaire, présente également des limites. La capacité d'identifier de manière statistiquement significative des marqueurs dont l'effet est faible nécessite d'avoir suffisamment de données pour pouvoir observer l'allèle alternatif chez suffisamment d'individus. Cela limite donc l'étude à des polymorphismes ayant une fréquence allélique minimale, généralement supérieure à 1% voire 5%. Les études pangénomiques n'ont généralement pas la puissance statistique nécessaire pour détecter des associations avec des polymorphismes rares. De même, plus le nombre d'individus inclus dans l'étude est élevé, plus la puissance statistique l'est également. Or, génotyper un nombre très élevé de SNPs sur un grand nombre d'individus représente un investissement considérable.

Les associations détectées par une approche pangénomique sont, comme toutes les études d'association, soumises à l'influence de potentiels facteurs confondants. Les principaux facteurs confondants pouvant biaiser les analyses sont l'existence d'une structure génétique dans la population qui différencierait les cas et les témoins, et un éventuel apparentement entre les individus inclus. L'appartenance à un groupe ethnique donné est généralement un des critères d'inclusion d'une étude pangénomique. Cependant, même au sein d'un groupe d'individus d'ascendance homogène tel que les Caucasiens d'origine européenne, il existe des différences subtiles mais fréquentes au sein du génome au niveau de certains polymorphismes. Or si les cas et les témoins ne présentent pas la même distribution au niveau de ces polymorphismes, les analyses statistiques concluront à une association entre ces polymorphismes et la maladie, alors qu'elles ne sont dues qu'à la constitution des deux groupes d'individus étudiés. Prenons l'exemple théorique d'une étude s'intéressant à la prédisposition génétique au cancer du poumon. Si on ne contrôle pas l'origine ethnique des individus inclus, il se peut que, par l'influence de facteurs externes (efficacité du recrutement dans certains centres, méthodes de recrutement, etc...), le groupe des cas soient enrichis en individus originaires d'Espagne, alors que le groupe des témoins sera enrichi en individus d'origine norvégienne. Les analyses statistiques concluront alors à tort que les polymorphismes situés sur le gène *MC1R*, gène codant pour la couleur des cheveux, sont associés avec le risque de cancer du poumon. Une structure de population basée sur un élément aussi frappant que la couleur des yeux ou des cheveux est relativement facile à détecter, mais toutes sont loin d'être évidentes.

Pour s'affranchir de toute stratification de population à l'échelle du génome entier, on utilise une méthode appelée analyse en composantes principales ou ACP, qui permet de maximiser la variabilité des données observées suivant les variables inclues dans l'étude. Les données sont projetées dans un repère dont les axes représentent une combinaison linéaire des variables, des SNPs dans notre cas. On représente généralement les résultats de l'ACP sur un graphe en deux dimensions avec les premiers axes de l'ACP en abscisses et en ordonnées. La figure 18 illustre le résultat de l'analyse en composantes principales réalisée sur 1387 individus européens sains appartenant au projet POPRES[115], dont les origines géographiques sont connues avec fiabilité. Ces individus ont été génotypés sur une puce SNP Affymetrix 500k.

Figure 18 – Projection des données génétiques de 1387 individus sur les deux premiers axes de l'ACP



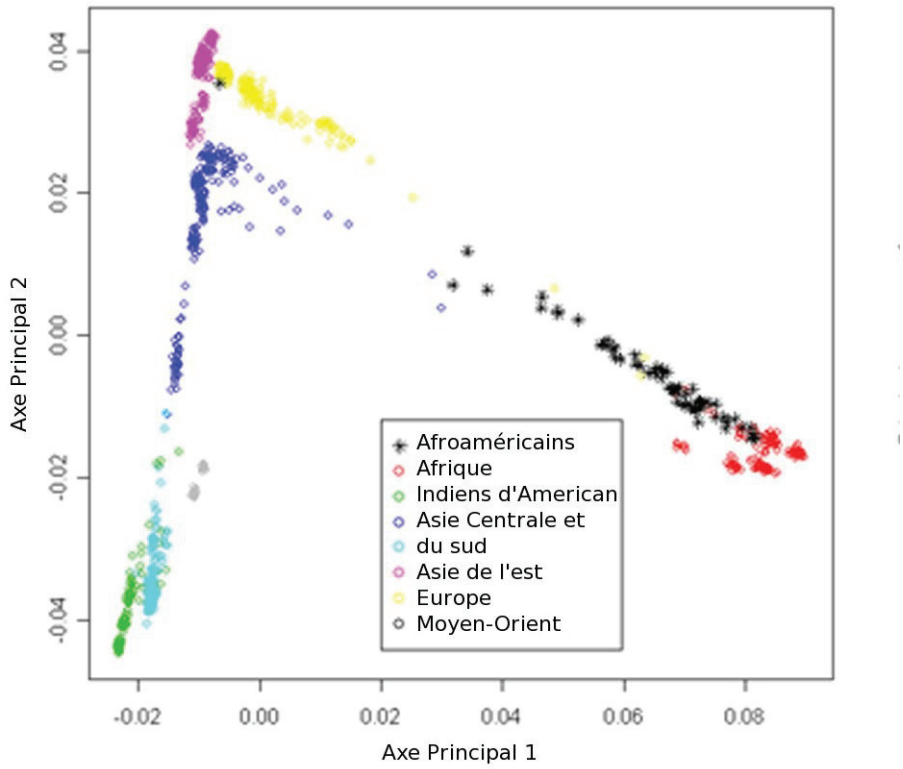
ACP réalisée sur 1387 individus sains dont les origines géographiques sont connues avec fiabilité, génotypés sur une puce Affymetrix 500k, appartenant à l'étude *Population Reference Sample*, ou POPRES[115]. PC1 et PC2 représentent les deux premiers axes de l'ACP effectuée. Chaque label coloré représente l'origine géographique d'un individu. Les pastilles colorées représentent la médiane des coordonnées des individus pour chaque pays. L'indicatif de chaque pays est représenté sur la carte dans le pays correspondant. Une rotation a été appliquée sur les axes pour souligner la similitude entre le deux représentations graphiques. *Novembre et al. Nature. 2008*[116]

On remarque une grande similitude entre la projection des individus sur les axes de l'ACP et l'origine géographique des individus. En construisant ce type de graphe, on peut ainsi vérifier

l'homogénéité génétique d'une population. Tous les individus inclus sont-ils Caucasiens ? Y a-t'il des individus qui se distinguent en terme de structure génétique ? Il est très difficile d'obtenir une population composée d'un nombre élevé d'individus et qui soit génétiquement homogène. Ainsi, afin de ne pas exclure inutilement des individus de l'étude, tout en tenant compte de la potentielle hétérogénéité génétique, on ajuste les modèles d'association construits en introduisant les coordonnées des individus sur les premiers axes de l'ACP, comme covariables dans le modèle.

Dans ce premier exemple, l'ACP est réalisée uniquement sur des individus européens. Seules quelques dizaines de SNPs discriminent les individus selon leur origine géographique en Europe. Mais l'ACP peut également être réalisée à l'échelle des populations, en discriminant les individus d'origine africaine, asiatique et européenne par exemple. Ce sont alors des portions entières de chromosomes qui sont utilisées pour distinguer ces populations. À cette échelle, la variabilité au sein d'une population est négligeable en comparaison des différences observées avec d'autres populations, ceci est illustré sur la figure 19, sur laquelle les populations d'origine africaine, européenne, et asiatique sont nettement séparées selon les deux premiers axes de l'ACP.

Figure 19 – Projection des données génétiques de 1087 individus sur les deux premiers axes de l'ACP



Représentation graphique des individus dans le plan des deux premiers axes de l'ACP réalisée. Chaque population et sous-population est représentée par une couleur différente. Analyse réalisée dans le cadre d'une étude du biais introduit dans les études GWAS lié à la stratification des populations auquelles appartiennent les individus génotypés. *adapté d'après Hao et al, PloS ONE. 2010*[117]

Afin de ne pas surestimer les fréquences alléliques dans le groupe des cas ou des témoins, les individus inclus dans une étude pangénomique ne doivent pas être apparentés, car leur données génétiques ne sont pas indépendantes. Les inclure biaiserait l'analyse statistique réalisée. De même, les SNPs conservés dans l'étude ne doivent pas violer de manière flagrante l'équilibre de Hardy-Weinberg. L'équilibre de Hardy-Weinberg est l'état d'une population dans lequel on peut prédire exactement les fréquences génotypiques à partir des fréquences alléliques. Soit un SNP biallélique a/A, dont les fréquences alléliques sont respectivement p et $q = 1-p$. Si ce SNP est à l'équilibre de Hardy-Weinberg, alors : $f(aa) = p^2$, $f(aA) = 2pq$ et $f(AA) = q^2$. L'équilibre de Hardy-Weinberg est une conséquence de l'appariement aléatoire des partenaires dans une population en l'absence de mutation, migration, sélection naturelle, ou dérive génétique. Concrètement, c'est l'état dans lequel les allèles hérités maternel et paternel d'un individu sont génétiquement indépendants. Cet équilibre n'est que très rarement parfaitement respecté en réalité, notre génome étant soumis à différentes forces évolutives qui influent sur les fréquences génotypiques. Par exemple, dans le cadre d'une maladie monogénique récessive à pénétrance complète affectant les enfants dès leur naissance et dont l'espérance de vie ne dépasserait pas une quinzaine d'années, ces individus n'auront très vraisemblablement pas de descendants, et leur génotype ne sera pas transmis, selon le principe de la sélection naturelle. Dans une étude pangénomique, il est toléré de conserver les SNPs qui ne s'éloignent que faiblement de l'équilibre de Hardy-Weinberg. Cependant, le fait que certains s'en éloignent trop peut indiquer un appariement non aléatoire des gamètes une possible stratification de population, des erreurs de génotypage non aléatoires, ou des données manquantes dans lesquelles un génotype est plus souvent absent que les autres. D'autre part, il a été estimé que le test de Hardy-Weinberg n'est pas très sensible aux erreurs de génotypage[118]. Afin d'exclure toute possibilité de biais, ces SNPs ne sont généralement pas inclus dans les analyses.

Toutes les pathologies ne se prêtent pas aux études pangénomiques. La nécessité d'avoir à disposition un nombre élevé de cas et de témoins peut poser problème dans la phase de recrutement, en particulier pour les pathologies les plus rares. De plus, une maladie qui se caractérise par des degrés d'atteintes multiples est souvent multifactorielle, et déterminer un phénotype homogène parmi les cas peut être limitant. Enfin, les interactions des facteurs génétiques avec des facteurs environnementaux peut complexifier la détection des associations.

Dans le cadre du cancer du sein, la très grande majorité des SNPs identifiés comme loci de prédisposition l'ont été grâce à des approches GWAS.

II.5 Le statut particulier du génome mitochondrial

Les méthodes de détection d'association présentées ici sont dédiées à l'identification de loci portés par le génome nucléaire humain. Cependant, la médecine actuelle doit maintenant prendre en compte l'influence « des autres génomes », ceux des entités avec lesquelles notre corps interagit directement. Il est de nos jours clairement établi que certaines souches virales peuvent favoriser le développement de cancers. Par exemple, l'infection par le virus de l'hépatite B et C favorise le développement de carcinome hépatocellulaire (cancer du foie)[119], mais certains polymorphismes à la fois dans le génome nucléaire et dans les gènes viraux modifient cette association. De même, le papillomavirus prédispose fortement au cancer du col de l'utérus[120]. Les virus sont les micro-organismes carcinogènes les plus fréquents, mais certains parasites et bactéries sont également classés comme tels[121]. Ainsi, la variabilité génomique observée au sein du viriome et du microbiome commence à être étudiée en tant que potentiel modificateur. De notre point de vue, il est tout aussi intéressant de s'intéresser à la variabilité du chondriome, soit l'ensemble des mitochondries de nos cellules. Bien qu'existant en de très nombreuses copies dans la très grande majorité de nos cellules, les mitochondries, selon l'hypothèse la plus probable, sont d'anciennes bactéries qui ont développé avec les cellules eucaryotes une association endosymbiotique, et qui se sont à terme entièrement intégrées dans nos cellules. De leur origine bactérienne, elles ont cependant conservé leur génome, qui est indépendant du génome nucléaire, et possède ses propres caractéristiques.

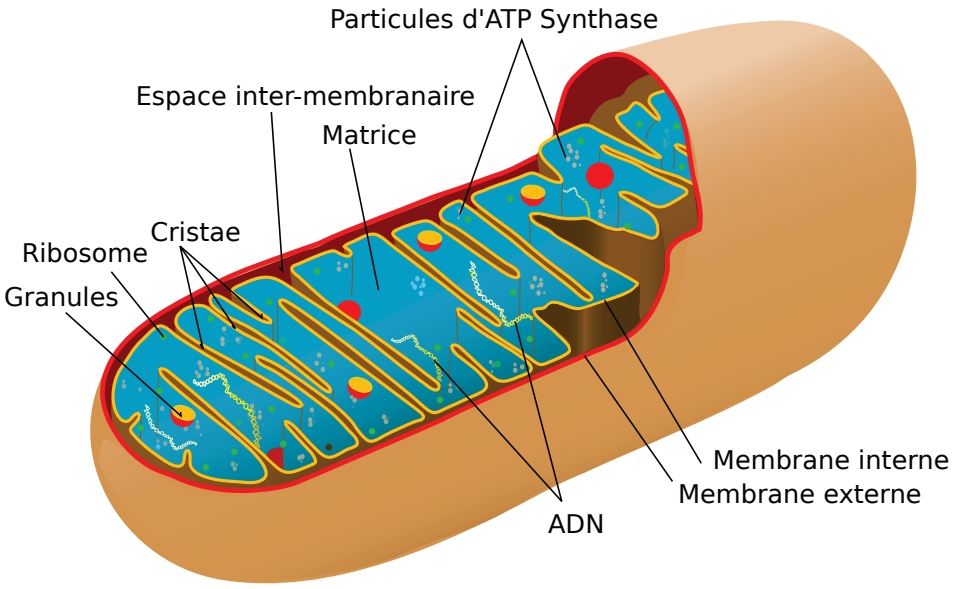
Le génome mitochondrial est distinct du génome nucléaire, et est porté par un chromosome distinct des 22 paires de chromosomes autosomiques et des 2 chromosomes sexuels. Les méthodes de détection d'association génomiques ne sont pas du tout optimisées pour étudier cette région. En effet, les analyses de liaison familiales se basent sur le mécanisme de recombinaison et sur l'analyse des blocs de déséquilibre de liaison. Hors le génome mitochondrial n'est pas sujet à la recombinaison. D'autre part, les puces commerciales nécessaires au génotypage pour les études pangénomiques ciblent effectivement certains polymorphismes mitochondriaux. Cependant, ils ne représentent qu'une faible proportion de l'ensemble des polymorphismes connus au sein du génome mitochondrial. De plus, le principe de détection indirecte d'association sur lequel est basée l'analyse des études pangénomiques n'est pas utile en l'absence de recombinaison. Du fait des limites de ces deux approches dans la détection d'associations génomiques au niveau du génome mitochondrial, il est légitime de s'interroger sur la possibilité que le génome mitochondrial soit sous-exploré dans la majorité des études de recherches d'association, ce qui en fait en soi une cible d'étude intéressante. L'étude de cette région génomique nécessite donc une approche ciblée plutôt que pangénomique, que ce soit par génotypage, ou par séquençage. De plus, la mitochondrie est une structure indispensable au fonctionnement de nos cellules, et assume des fonctions métaboliques essentielles. L'ensemble de ces considérations justifient l'étude de la variabilité de son génome dans le contexte du cancer, et dans notre cas dans le cadre du cancer du sein.

III. La mitochondrie, un organite essentiel ayant sa propre histoire

III.1 Structure de l'organite et de son génome

La mitochondrie est un organite présent dans la majorité des cellules de notre corps. On les trouve en un nombre variable d'exemplaires dans nos cellules, allant de quelques copies dans les plaquettes sanguines, quelques centaines dans les cellules de la peau, plusieurs milliers de copies dans les cellules cardiaques, hépatiques, et cérébrales, et jusqu'à plusieurs dizaines de milliers dans les oocytes[122]. Les mitochondries possèdent deux membranes, une interne et une externe, définissant un espace intermembranaire (Figure 20). La membrane interne délimite la matrice mitochondriale, qui contient notamment les ribosomes mitochondriaux, des granules contenant du phosphate de calcium, et l'ADN mitochondrial. La membrane interne se replie vers l'intérieur de la mitochondrie en formant des crêtes mitochondriales, appelées également *cristae*, augmentant ainsi la surface totale de la membrane interne. Au niveau des crêtes se concentrent un grand nombre de protéines membranaires.

Figure 20 – Représentation schématique et principaux composants d'une mitochondrie



Le diamètre des mitochondries est de l'ordre de quelques micromètres, mais leur forme et leurs dimensions sont extrêmement variables en fonction des tissus observés. Les mitochondries des cellules musculaires sont caractérisées par une forme en filament et sont très allongées en comparaison des mitochondries baignant plus librement dans le cytoplasme de cellules subissant moins de contraintes, qui auront une forme plus classique sphérique ou ellipsoïdale. De plus, des études ont formulé l'hypothèse que la morphologie des mitochondries pourrait être une conséquence de leur rôle fonctionnel dans nos cellules[123–125]. En effet, les mitochondries sont des composants dynamiques au sein de nos cellules. Elles ont aussi bien la capacité de se scinder

en plusieurs copies que de fusionner avec d'autres, et ce en fonction des besoins énergétiques de la cellule et des marqueurs de signalisation qu'elles reçoivent[126], de manière complètement indépendante du cycle de division de la cellule dans laquelle elles se situent.

La mitochondrie possède son propre génome, qu'on désigne de manière générale par le chromosome M ou MT. L'ADN mitochondrial - ou ADNmt - se présente sous la forme d'un fragment d'ADN circulaire et haploïde (Figure 21). La plupart du temps, il existe plusieurs copies de cet ADN dans une mitochondrie, notamment à cause de la dynamique de fusion/fission à laquelle elles sont soumises. Le génome mitochondrial sous sa forme la plus fréquente mesure 16 569 bases (environ 17kb). Il porte 37 gènes, qui ne comportent pas d'introns. La mitochondrie n'utilise pas les mécanismes de synthèse protéique de sa cellule-hôte, mais possède son propre système de synthèse protéique. C'est pourquoi 2 des 37 gènes portés par le génome mitochondrial sont des ARNs ribosomiques (ARNr), respectivement nommés *MT-RNR1* et *MT-RNR2*. Ils codent pour les ARNr 12S et 16S qui sont des composants structuraux essentiels des ribosomes mitochondriaux, appelés également mitoribosomes. De plus, la mitochondrie utilise également ses propres ARNs de transfert (ARNt) pour la synthèse des acides aminés qu'elle utilise. On trouve donc 22 gènes codant pour ces ARNt. Enfin, la mitochondrie n'utilise pas le même code génétique que celui utilisé par le reste de l'organisme humain. Les 13 gènes restants sont des gènes codants pour des protéines membranaires mitochondriales ayant un rôle structurel dans les complexes de la phosphorylation oxydative (voir Section III.2).

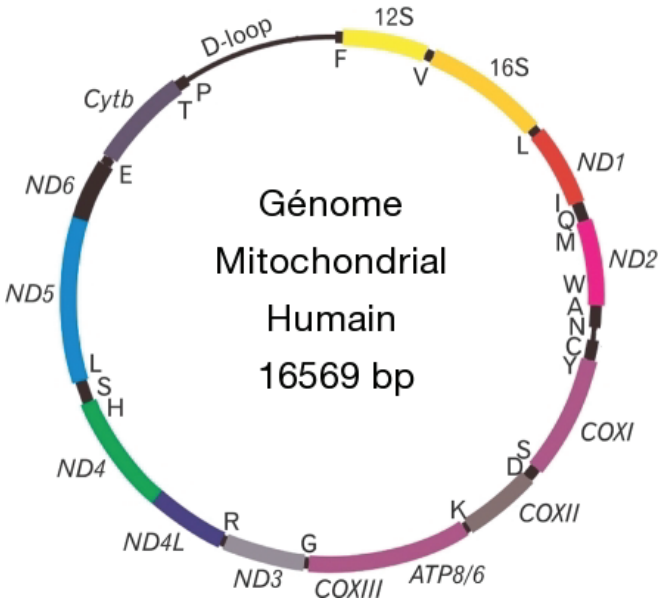


Figure 21 – Structure et gènes portés par le génome mitochondrial

Les gènes codant pour des protéines sont représentés en italique. Les gènes codant pour les ARNs de transfert mitochondriaux sont représentés par le symbole de l'acide aminé auquel ils correspondent. Les deux gènes ribosomiques *MT-RNR1* et *MT-RNR2* sont représentés par le nom de l'ARN ribosome qu'ils encodent, respectivement 12S et 16S. *Yoon et al. Anat. Cell Biol. 2010*[127]

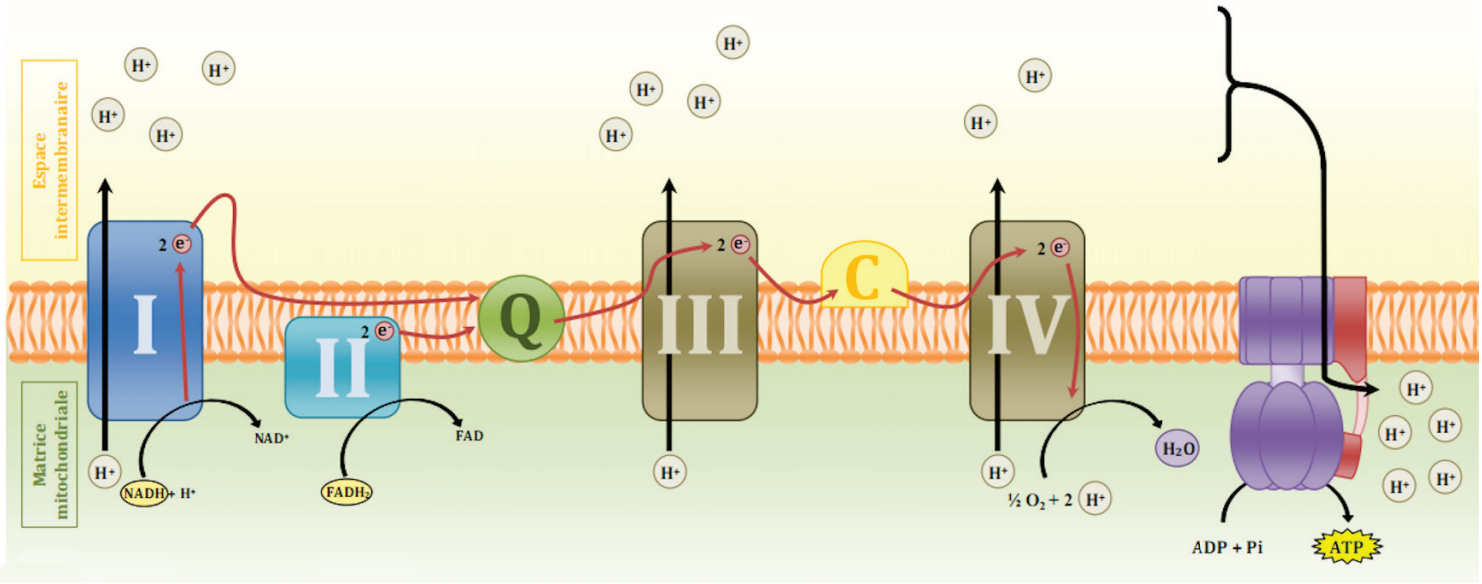
Il est usuellement admis que le génome mitochondrial ne subit pas de processus de recombinaison. Dans le cas de la mitochondrie, l'absence de ce mécanisme signifie que les mitochondries ne peuvent pas échanger des fragments d'ADNmt entre elles, et ne peuvent pas conduire à la formation de mitochondries génétiquement hybrides. De plus, le génome mitochondrial est hérité par la lignée maternelle[128]. En effet, au moment de la fécondation, à la fois les mitochondries provenant de l'ovule maternel et celles provenant du spermatozoïde paternel sont présentes dans la cellule-oeuf. Cependant, les mitochondries paternelles sont aussitôt marquées à l'ubiquitine, et détruites[129]. En l'absence de recombinaison, aucun échange entre les mitochondries maternelles et paternelles n'est possible. Des études ont cependant mis en évidence dans de très rares cas une potentielle transmission de l'ADNmt paternel à la descendance[130]. Une autre est allée plus loin et a conclu avoir mis en évidence des fragments d'ADN mitochondrial issu de la recombinaison entre le génome mitochondrial paternel et maternel dans des cellules somatiques[131]. Ces études ouvrent de nouvelles perspectives sur la transmission du génome mitochondrial, mais dans tous les cas les phénomènes associés semblent se produire à une fréquence très faible, ce qui les rend difficilement perceptibles[132,133].

Un individu peut avoir au sein de certaines cellules d'un de ses organes des mitochondries ayant un génome mitochondrial différent de celui porté par la majorité de ses mitochondries. Cet état est appelé hétéroplasmie. L'hétéroplasmie est la conséquence de l'apparition d'une mutation au sein du génome mitochondrial, qui est propagée lors de la fission des mitochondries à plusieurs, voir de nombreuses copies de la mitochondrie dans laquelle la mutation est initialement apparue. Cet état hétéroplasmique est généralement perpétué lors de la division cellulaire, en fonction de la répartition des mitochondries de la cellule en division entres ses deux cellules filles. L'hétéroplasmie peut alors atteindre une fréquence importante au sein d'un organe donné. L'hétéroplasmie peut également être héritée si l'oocyte fécondé lors de la conception était également hétéroplasmique[134]. L'hétéroplasmie peut n'avoir aucune manifestation phénotypique si la mutation propagée n'a pas de conséquence pathologique. Ainsi, Ramos *et al.*[135] ont étudié la fréquence et la distribution des variants hétéroplasmiques du génome mitochondrial sur un échantillon sanguin chez 101 individus espagnols sains. La méthode utilisée permet de détecter une hétéroplasmie ayant une fréquence supérieure à 10% de l'échantillon. Les conclusions de cette étude ont montré qu'environ 61% des individus analysés présentaient un variant hétéroplasmique, dont 24% au moins une hétéroplasmie ponctuelle (de type substitution), et 49 % une hétéoplasmie liée à une variation du nombre de répétitions d'une courte séquence du type microsatellite. Cependant, une mutation pathogène propagée de manière hétéroplasmique peut avoir de graves conséquences sur l'organisme (voir Section III.4). La degré de gravité des maladies causées par une mutation hétéroplasmique est souvent lié à la fréquence d'hétéroplasmie de cette mutation dans un organe, notamment quand cette fréquence dépasse un seuil au delà duquel certains symptômes se manifestent[136,137].

III.2 Fonctions de la mitochondrie

Le mitochondrie assume plusieurs rôles au sein de nos cellules. La principale fonction de la mitochondrie est sans aucun doute la synthèse de l'Adénosine TriPhosphate, ou ATP, l'énergie dont nos cellules ont besoin pour fonctionner. L'ATP est utilisé dans tout un ensemble de processus de la cellule, tels que le transport actif d'ions à travers les membranes, ou par les cellules musculaires lors de la transformation d'énergie chimique en énergie mécanique. La synthèse d'ATP est réalisée par un ensemble de réactions chimiques appelée phosphorylation oxydative (Figure 22).

Figure 22 – Mécanisme de synthèse de l'ATP par phosphorylation oxydative au niveau de la membrane interne de la mitochondrie.



Le transfert d'électrons de la matrice mitochondriale à l'espace intermembranaire au niveau des complexes protéiques formant la chaîne respiratoire engendre un gradient de protons H+. La phosphorylation oxydative se conclut par la synthèse d'ATP par l'ATP-synthétase, une pompe qui utilise l'énergie des protons qu'elle transfère de l'espace intermembranaire à la matrice mitochondriale pour phosphoryler l'ADP en ATP, réaction impliquant également un Phosphate inorganique. *Complexe I* :NADH coenzyme réductase. *Complexe II* : Succinate coenzyme Q réductase. *Complexe III* : Coenzyme Q cytochrome C réductase. *Complexe IV* : Cytochrome C oxydase. *Q* : Coenzyme Q ou ubiquinone. *C* : Cytochrome C.

Au cours de ce processus, des électrons sont transférés d'espèces chimiques réductrices NADH (NADH $\longrightarrow$ $NAD^+ + H^+ + 2\,e^-$, avec NAD^+ : nicotinamide adénine dinucléotide) et $FADH_2$ (FADH $\longrightarrow$ $FAD + 2\,H^+ + 2\,e^-$ avec FAD : flavine adénine dinucléotide) vers des espèces oxydantes par oxydoréduction, générant ainsi une gradient de protons H+. Ce transfert d'électrons est effectué par un ensemble de quatre complexes protéiques intégrés à la membrane interne mitochondriale, formant la chaîne respiratoire. Le complexe I a une action NADH coenzyme réductase, récupérant les électrons du NADH et permettant le transport de 4 protons de

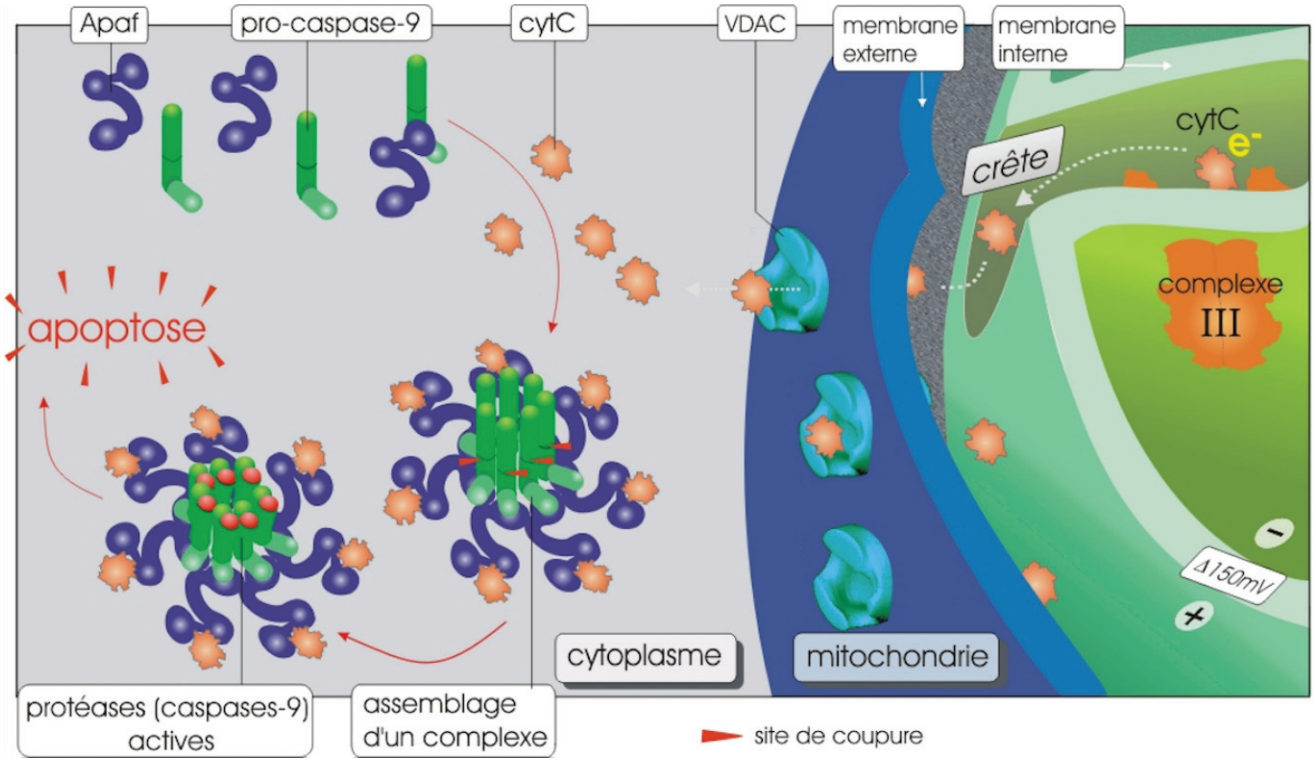
la matrice mitochondriale à l'espace intermembranaire. Le complexe II a une action succinate coenzyme Q réductase, récupérant les électrons du $FADH_2$, et ne transportant aucun proton. Le complexe III a une action coenzyme Q cytochrome C réductase, et permet le transport de 4 protons de la matrice mitochondriale à l'espace intermembranaire. Le complexe IV a une action cytochrome C oxydase, et permet le transport de 2 protons de la matrice mitochondriale à l'espace intermembranaire. Le coenzyme Q (ou ubiquinone) permet la transition entre le complexe I/II et le complexe III. Le cytochrome C permet la transition entre le complexe III et le complexe IV. La majorité des protéines structurelles de la chaîne de phosphorylation oxydative sont encodées par le génome nucléaire, à l'exception de certaines de ses sous-unités : sept sous-unités du complexe I, une du complexe II, une du complexe III, trois du complexe IV, et deux du complexe V.

Suite à la chaîne de complexes protéiques, le dernier accepteur d'électrons est l'oxygène qui sera ainsi à l'origine de la formation d'une molécule d'eau. Le NADH permettra donc le transport de 10 protons de la matrice mitochondriale à l'espace intermembranaire, tandis que le $FADH_2$ en permettra le transport de seulement 6. Ceux-ci repassent vers la matrice mitochondriale via une pompe à protons appelée ATP-synthétase, et qui produit l'ATP à partir d'ADP (Adénosine DiPhosphate) et de Phosphate inorganique Pi.

La mitochondrie a également pour fonction de participer à la synthèse des hormones stéroïdes générées à partir du cholestérol, par l'intermédiaire du cytochrome P450. En fonction de l'organe dans lequel elle est effectuée, cette biosynthèse peut aboutir à la formation de testostérone (testicules), d'oestradiol et de progestérone (ovaires), ou de glucocorticoïdes tels que le cortisol (glandes surrénales). D'autre part, avec le réticulum endoplasmique, la mitochondrie est le principal réservoir cellulaire de calcium. Elle est impliquée dans la régulation de sa concentration dans le cytoplasme.

Enfin, la mitochondrie est impliquée dans le processus de la mort cellulaire programmée, l'apoptose (Figure 23). L'activation de ce processus peut être provoquée par divers mécanismes, notamment en cas de stress cellulaire ou de hausse de la concentration de Ca^{2+}[138,139]. Elle se manifeste par une hausse de la perméabilité membranaire mitochondriale due à l'activation des pores de transition de perméabilité mitochondriale - ou MPT pour *Mitochondrial Permeability Transition pore* - situés sur la membrane interne de la mitochondrie. L'activation des pores mitochondriaux provoque l'augmentation de la perméabilité mitochondriale à des molécules de poids moléculaire inférieur à 1500 Da. L'augmentation de la perméabilité de la membrane mitochondriale s'accompagne d'une baisse du gradient électrochimique nécessaire à la production d'ATP, et diminue les capacités de synthèse énergétique de la mitochondrie[140]. De plus, l'activation des pores mitochondriaux entraîne le relargage dans le cytoplasme du cytochrome C. Le cytochrome C participe alors à l'assemblage de l'apoptosome, un complexe protéique impliquant les protéines Apaf et pro-caspase 9. Ce complexe est nécessaire afin d'activer d'autres protéines caspases, protéines qui détruisent de nombreux composants moléculaires du noyau et du cytoplasme, conduisant à la mort de la cellule.

Figure 23 – Initiation de l'apoptose par relargage du cytochrome C mitochondrial



L'apoptose peut être initiée par la mitochondrie en relarguant le cytochrome C dans le cytosol, par l'intermédiaire des canaux VDAC (Voltage-Dependant Anion Channel). Le cytochrome C est nécessaire à l'assemblage du complexe apoptotique impliquant les protéines Apaf et pro-caspase-9. Une fois l'apoptosome assemblé, il active à sont tour d'autres caspases, responsables de la dégradation de composants essentiels du noyau et du cytoplasme, induisant la mort cellulaire.

III.3 Evolution du génome mitochondrial : la notion d'haplogroupe

Comme mentionné précédemment, le génome mitochondrial est haploïde, ne subit pas de recombinaison et est transmis uniquement par la lignée maternelle. Ces caractéristiques en font un modèle d'étude de l'évolution des génomes des plus intéressants.

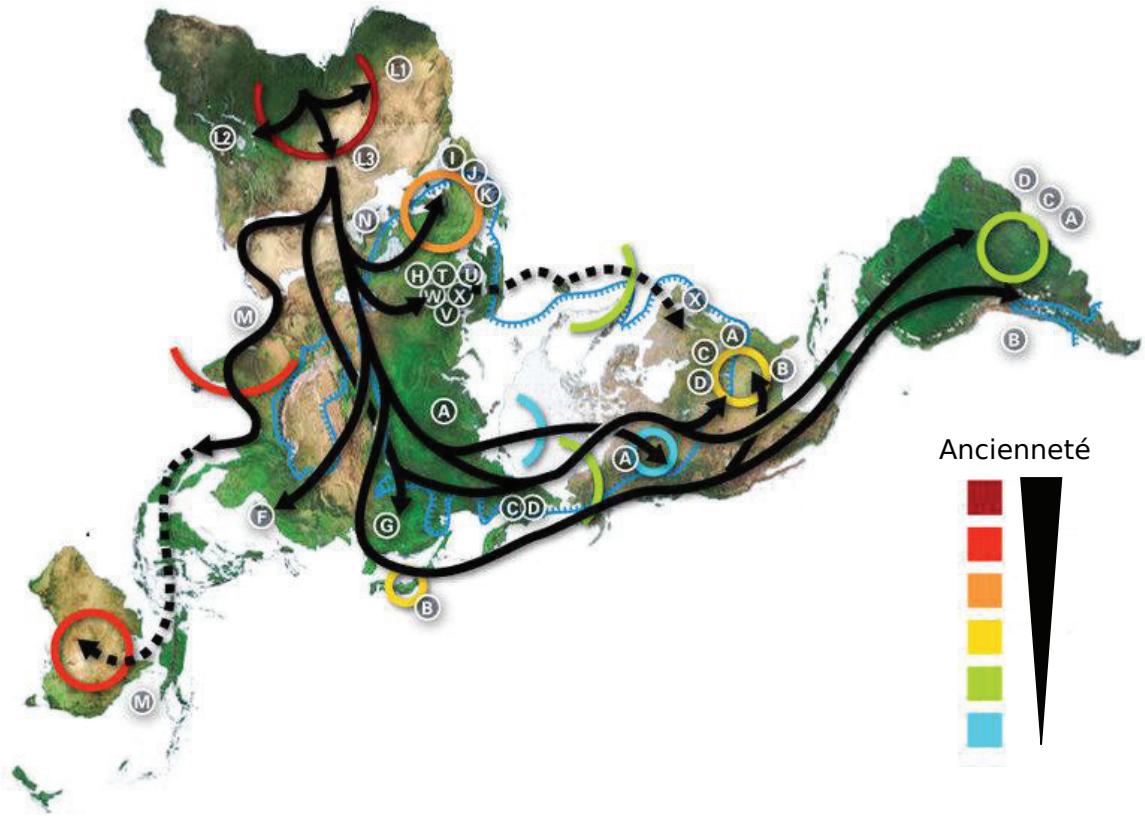
Il a été établi que le taux de mutation intrinsèque du génome mitochondrial est beaucoup plus élevé que celui du génome nucléaire[141] avec une valeur de 10 à 200 fois supérieure, selon les estimationss[142]. Plusieurs facteurs expliquent ce taux de mutation élevé : l'absence de protection de l'ADN par des histones comme c'est le cas pour le génome nucléaire, un mécanisme de correction des erreurs d'incorporation de nucléotides par les polymérases (mécanisme appelé *proof-reading*) dont l'efficacité est moindre que le mécanisme s'exerçant dans le génome nucléaire[143], et enfin l'exposition directe de l'ADN mitochondrial à des composés occasionnant des dommages à l'ADN.

Le taux de mutation du génome mitochondrial étant élevé, des mutations apparues de manière spontanée chez un individu ont été transmises à sa descendance, et de nombreux polymorphismes ont été observés sur les quelques 17000 paires de bases qui constituent ce génome. Comme présenté en Section II.1 , on appelle haplotype une combinaison d'allèles localisés sur différents loci d'un même chromosome. On peut donc caractériser à la fois la variabilité du génome mitochondrial en termes de polymorphismes, mais également la manière dont ces polymorphismes sont combinés entre eux en étudiant les haplotypes du génome mitochondrial.

L'évolution du génome mitochondrial a eu lieu en parallèle de l'expansion des premiers *Homo sapiens* hors d'Afrique[144]. Certaines des mutations acquises sont donc devenues caractéristiques de groupes ethniques existants sur la planète. Ces mutations sont devenues des critères de regroupement des haplotypes mitochondriaux en *haplogroupes* mitochondriaux. Tous les haplotypes mitochondriaux regroupés sous un même haplogroupe sont dérivés du même ancêtre commun. Ces haplogroupes mitochondriaux sont désignés par des lettres de l'alphabet, actuellement toutes utilisées à l'exception de la lettre O. Ainsi, les principaux haplogroupes mitochondriaux observés dans la population européenne sont les haplogroupes H, I, J, K, T, U, V, W, et X. Les haplogroupes asiatiques sont A, B, C, D, F, G, M, Y, et Z. Les haplogroupes africains sont principalement représentés par l'haplogroupe L et ses sous-clades. Les haplogroupes observés actuellement sur le continent américain sont des haplogroupes européens chez les blancs non-hispaniques, africains chez les noirs non-hispaniques, et d'origine asiatique chez les Mexicains et les autres Hispaniques[145]. Les ethnies vivant sur le continent américain avant la colonisation par les européens telles que les esquimaux, différentes tribus indiennes, ou les aztèques sont toutes caractérisées par des haplogroupes d'origine asiatique également[146]. De plus, certains groupes ethniques spécifiques présentent un haplogroupe propre. Par exemple, l'haplogroupe des Inuits vivant au nord de l'Alaska et dans le grand nord canadien est A2[146].

Sur l'échelle de l'Évolution, les haplogroupes les plus anciens sont les haplogroupes originaires d'Afrique, berceau de l'humanité (Figure 24). Les humains ont ensuite gagné le Moyen-Orient, puis ont colonisé l'Europe, et enfin l'Asie. Les hypothèses sur la colonisation de l'Amérique divergent encore, mais il semblerait que plusieurs flux migratoires successifs soient partis de différents points d'Asie pour aller coloniser l'Amérique du nord et du sud[146]. L'analyse des données du génome mitochondrial a contribué à valider, réfuter, ou nuancer les hypothèses existantes sur les flux migratoires humains.

Figure 24 – Représentation simplifiée des flux migratoires en fonction des haplogroupes mitochondriaux



La forte variabilité du génome mitochondrial, sa structure haploïde, et le fait qu'il ne soit transmis que par la lignée maternelle sans recombinaison ont permis de reconstruire son évolution depuis l'expansion des premiers hommes hors d'Afrique. L'arbre phylogénétique du génome mitochondrial a été reconstitué à partir de l'ensemble des séquences disponibles dans les bases de données publiques[147]. Cet arbre, appelé PhyloTree[148], est actuellement l'arbre phylogénétique de référence pour le génome mitochondrial. Il formalise la manière dont les différents haplogroupes ainsi que toutes leurs sous-clades ont évolué les uns par rapport aux autres. Chaque sous clade est définie par un ensemble de mutations dont les allèles sont précisés. Il est accessible via l'url `http://www.phylotree.org/`, et est régulièrement mis à jour pour prendre en compte l'ensemble des séquences disponibles publiquement.

La séquence du génome mitochondrial a été publiée pour la première fois en 1981[149], sous le nom de *Cambridge Reference Sequence*, ou CRS. Les principaux gènes mitochondriaux ont été localisés ou prédits de manière correcte sur CRS. En 1999, une version révisée de la séquence CRS, appelée rCRS pour *revised Cambridge Refarence Sequence* a été publiée[150] à la suite de l'identification d'erreurs au niveau de 11 positions nucléotidiques. En particulier, la position n° 3107 dans la séquence originale CRS avait été introduite à tort. Afin de ne pas perturber les habitudes de numérotation des bases sur le génome mitochondrial, et ne pas introduire de décalage suite à cette correction, la numérotation originale des nucléotides est conservée dans rCRS ; la position n° 3107 est représentée par un « N » dans la séquence de référence sur Genbank (base de données de référence des séquences nucléotidiques). La publication de rCRS a également permis de confirmer que la séquence initiale CRS porte l'allèle rare de sept polymorphismes, et ne sont pas des erreurs. Cette séquence rCRS est accessible sur Genbank sous le numéro d'accession `NC_012920.1`. rCRS est la séquence de référence du génome mitochondrial utilisée dans l'assemblage GRCh37 du génome complet humain.

L'information portée par le génome mitochondrial est utilisée dans de nombreux champs disciplinaires. En médecine génétique, le génome mitochondrial est analysé dans le cadre du diagnostic ou de la recherche de variants de prédisposition de certaines pathologies. En sciences forensiques, il est utilisé pour identifier ou établir des liens de parenté entre plusieurs individus. Mais comme évoqué plus haut, le génome mitochondrial est également très utilisé à la fois comme support pour étudier les flux de migrations humaines, et comme concept d'étude de l'évolution des génomes. Certains chercheurs de cette branche trouvaient qu'il était inapproprié d'utiliser comme séquence de référence rCRS, soir la séquence du génome mitochondrial d'un individu originaire du Royaume-Uni ayant vécu au vingtième siècle, et portant en certaines positions des allèles non-ancestraux. C'est pourquoi ils ont cherché à reconstituer la séquence du génome mitochondrial du plus récent ancêtre commun de tous les hommes. À partir de l'ensemble des séquences, ils ont reconstruit par parcimonie la séquence ancestrale du génome mitochondrial, appelée RSRS, pour *Reconstructed Sapiens Reference Sequence*[151]. Cette séquence est reconstruite *in silico*, et n'a pas été observée directement. Elle contient les allèles ancestraux des polymorphismes identifiés à ce jour.

MITOMAP[152,153] est une base de données accessible en ligne regroupant de très nombreuses informations sur la variabilité du génome mitochondrial. On y trouve la liste des polymorphismes et des mutations publiées ou non publiées, ainsi que les références associées, en fonction de leur localisation dans la région codante ou dans la région régulatrice du chromosome mitochondrial. On y trouve également les altérations plus importantes du génome (délétions majeures, réarrangements simples et complexes), la description de mutations somatiques, et des références sur les associations entre des variants du génome mitochondrial et des pathologies, ainsi que des informations sur l'implication des gènes nucléaires dans les atteintes mitochondriales.

III.4 Pathologies connues liées à des altérations génomiques mitochondriales

On dénombre à l'heure actuelle plusieurs centaines de pathologies liées à la mitochondrie dont les symptômes sont extrêmement variés. Une partie de ces pathologies a pour origine des altérations situées au niveau de l'ADN mitochondrial, alors que les autres sont dues à des altérations portées par des gènes nucléaires, mais ayant un impact sur le nombre, la structure, et les fonctions des mitochondries. Par exemple, des altérations du gène nucléaire *POLG*[154] se trouvent à l'origine de plusieurs pathologies dites mitochondriales. Ce gène code pour la seule ADN-polymérase effective dans la mitochondrie, la polymérase-γ[155], qui fait partie du complexe en charge de la réplication de l'ADN mitochondrial. Les altérations de ce gène ont été reportées dans des cas de syndrome adPEO[156] - *Autosomal Dominant chronic Progressive External Ophthalmoplegia* - mais ont également été associées à une incidence accrue de la maladie de Parkinson, à une ménopause précoce[157,158], ainsi qu'à la stérilité masculine[159,160]. Les autres pathologies impliquant la mitochondrie, mais trouvant leur origine au sein du génome nucléaire ne seront pas détaillées ici. Les pathologies dont la cause est une altération du génome mitochondrial sont généralement caractérisées par des atteintes ophtalmiques, neurologiques et nerveuses, musculaires, ou affectant l'ouïe[161,162].

Syndrome MELAS Le syndrome MELAS[163] - *Mitochondrial Encephalopathy, Lactic Acidosis, and Stroke-like syndrome* peut avoir plus d'une douzaine d'origine génétiques ponctuelles. Cette pathologie se manifeste par un développement cognitif limité et une atteinte de démence à divers degrés, un taux élevé de lactate dans le sang et dans le liquide cérébro-spinal, des accidents vasculaires cérébraux, des pseudo-épisodes vasculaires cérébraux, ainsi que des myopathies. Ce syndrome peut également être associé à une perte d'audition, du diabète, des migraines, des vomissements chroniques, problèmes de mobilité, et par le syndrome cardiaque de Wolff–Parkinson–White. Les mutations causales du syndrome de MELAS sont principalement localisées dans le gène codant pour l'ARN de transfert mitochondrial de la Leucine - plus précisément : A3243G (80%), T3271C (7 %), A3260G (environ 5 %), A3252G (< 5 %) - et dans d'autres gènes mitochondriaux, comme la mutation G13513A dans *ND5*.

Syndrome MERRF Le syndrome MERRF[164] ou épilepsie myoclonique avec fibres rouges déchiquetées - *Myoclonic Epilepsy and Ragged-Red Fibers* - est une encéphalomyopathie mitochondriale caractérisée par des crises myocloniques (contraction brèves et non rythmée des muscles). Elle est parfois associée à une surdité neurosensorielle, une atrophie optique, une petite taille ou une neuropathie périphérique. La maladie est progressive avec aggravation régulière de l'épilepsie, et apparition de symptômes additionnels comme l'ataxie (déficit de la coordination des muscles), une surdité, une faiblesse musculaire, ou une dégradation intellectuelle. Les causes génétiques principales sont les mutations dans le gène codant pour l'ARN de transfert mitochondrial Lysine dont A8344G (80 % des cas), et T8356C, G8363A, et G8361A représentant 10 % des cas. Des mutations dans les gènes codant pour les ARNs de transfert mitochondriaux Phénylalanine (G611A) et Proline (G15967A) sont également connues pour causer le syndrome

MERRF.

Syndrome NARP Le syndrome NARP - ou *Neuropathy Ataxia Retinitis Pigmentosa* se caractérise par une progressive apparition de neuropathies, d'ataxie, et de rétinite pigmentaire (dégénérescence progressive des cellules photosensibles de la rétine), accompagnée par le développement de démence. Ce syndrome est causé par la mutation T8993C lorsque celle-ci est hétéroplasmique et représente environ 70% à 90% des mitochondries d'un organe, notamment dans les tissus cérébraux[165,166].

Atrophie optique de Leber L'atrophie optique de Leber[167,168] ou LHON - *Leber hereditary optic neuropathy* - est caractérisée par une perte indolore et progressive de la vue, qui peut être associée à des maladies de type sclérose en plaques multiple. Plus de 90% des cas de LHON portent une des trois mutations primaires G3460A, G11778A, et T14484C. D'autres mutations causales plus rares sont répertoriées : G3635A, G3700A, G3733A, C4171A, T10663C, G13359A, C13382A, C13382G, A14495G, T14502C, C14568T

Syndrome Leigh Le syndrome de Leigh, ou encéphalomyopathie nécrosante subaiguë[169,170], se caractérise par un déclin progressif des fonctions neurologiques dues à l'apparition de lésions cérébrales. Les caractéristiques cliniques et génomiques de cette pathologie sont très hétérogènes. Les altérations génomiques à l'origine de ce syndrome sont majoritairement nucléaires, mais 10% à 30% des syndromes de Leigh sont dus à des mutations portées par l'ADN mitochondrial. Le gène mitochondrial le plus fréquemment muté dans ce syndrome est *ATP6* qui code pour une sous-unité de l'enzyme ATPase, ou complexe V de la phosphorylation oxydative, en charge de la synthèse d'ATP. Des altérations dans d'autres gènes, tels que les gènes *ND1* à *ND6*, *COX3*, ou dans ceux codant pour les ARNs de transfert mitochondriaux sont la cause du syndrome de Leigh.

Syndrome de Kearn-Sayre Le syndrome de Kearn-Sayre[171] ou KSS se caractérise principalement par une paralysie progressive des muscles moteurs des yeux (ophtalmoplégie) et par une rétinite pigmentaire (dégénérescence progressive des cellules photosensibles de la rétine). Les symptômes associés peuvent être la surdité, ainsi que des atteintes cardiaques, cérébrales musculaires, endocriniennes, et rénales. Ce syndrome a pour cause une large délétion au sein du génome mitochondrial, dont la taille la plus fréquemment observée est 4977 paires de bases (dans 30% des cas), et qui se manifeste de manière hétéroplasmique. La manifestation des symptômes dépend de la proportion des mitochondries portant cette délétion dans un tissu, qui est de 60% pour les muscles squelettiques.

IV. Mitochondrie, Stress oxydatif et Cancer

Il existe des pathologies qui trouvent leur origine dans des dysfonctionnements au niveau du génome mitochondrial. Mais qu'en est-il du rôle de la mitochondrie dans le cancer ? Les efforts soutenus de la recherche contre le cancer ont abouti à la définition de propriétés caractéristiques du développement tumoral[172]. Ces capacités biologiques acquises au cours de ce processus sont :

- l'activation de cascades de signalisation favorisant la prolifération cellulaire ;
- une déficience en facteurs d'inhibition de la croissance cellulaire ;
- la résistance à la mort cellulaire programmée (apoptose) ;
- l'activation de la réplication anarchique des cellules ;
- l'activation de l'angiogénèse (développement et croissance de nouveaux vaisseaux sanguins) ;
- l'acquisition des mécanismes d'invasion (colonisation de tissus différents au sein d'un même organe) et du potentiel métastatique (migration vers d'autres organes par le système sanguin ou lymphatique) ;
- la dérégulation énergétique, en favorisant la glycolyse en situation aérobie (lorsque suffisamment d'oxygène est à disposition), voie de synthèse énergétique bien moins efficace que la phosphorylation oxydative (effet Wargburg) ;
- l'insensibilité au contrôle exercé par le système immunitaire.

La mitochondrie est impliquée de manière directe dans au moins deux des mécanismes altérés lors du développement tumoral : l'induction de l'apoptose et la production énergétique de la cellule. Cependant, le rôle exact exercé par la mitochondrie dans le développement tumoral est aujourd'hui controversé. En particulier la question se pose de déterminer si la mitochondrie est impliquée dans l'apparition des premières néoplasies conduisant au cancer ou si les altérations de son génome et l'inactivation de certaines de ses fonctions ne sont qu'une conséquence de l'instabilité génomique tumorale[173].

IV.1 Altérations somatiques du génome mitochondrial dans la tumeur

Du fait de le l'activation d'oncogènes, de l'inactivation des gènes suppresseurs de tumeur, et des dysfonctionnements des mécanismes de réparation d l'ADN, le génome nucléaire des cellules tumorales se trouve drastiquement remanié dans la plupart des tumeurs. Les modifications du génome observées uniquement dans les cellules tumorales et pas dans les cellules normales sont dites somatiques. On peut ainsi se demander dans quelle mesure le génome mitochondrial est lui-même affecté par ses altérations dans les cellules tumorales. Les recherches menées depuis les deux dernières décennies ont permis d'identifier de nombreuses mutations somatiques du génome mitochondrial dans de nombreux types tumoraux, que ce soit des tumeurs solides ou des cancers hématologiques comme des leucémies ou des lymphomes[142].

À partir d'analyses effectuées sur un panel d'études récentes comprenant plus de 800 patients diagnostiqués pour 23 types de cancers différents, Yu *et al.*[142] ont montré que 57,7 % des individus étudiés présentaient des mutations ponctuelles au sein du génome mitochondrial de leur échantillon tumoral. Parmi ces mutations, environ 38% étaient situées au niveau de la région de contrôle (D-loop) du génome mitochondrial, 13% au niveau des gènes codant pour les ARNs ribosomiques mitochondriaux, et environ 50% dans les gènes codant pour des sous-unités des complexes formant la chaîne respiratoire mitochondriale. Une large proportion des mutations détectées situées dans des régions non conservées du génome mitochondrial est donc considérée comme non critique car n'affectant pas la séquence en acides aminés. À l'inverse, les 25% restants affectent directement la séquence codante, et les mutations non-synonymes (celles induisant des décalages de phase de lecture de l'ADN et celles introduisant ou supprimant des codons de terminaison) étaient principalement détectées dans des régions conservées du génome mitochondrial, pouvant ainsi altérer les fonctions des gènes localisés sur ces régions.

Des délétions de petite taille (moins de 300 bp) ont été observées de manière relativement sporadique dans divers cancers : cancer colorectal[174], de la rate[175], de l'estomac[176], et hépatocarcinome cellulaire[123,177]. Une délétion beaucoup plus large de 4977 paires de bases a été observée fréquemment, et ce dans de nombreux types de cancers : cancer du sein[178,179], cancer de l'endomètre[180], de l'oesophage[181,182], de l'estomac[183,179], hépatocarcinome cellulaire[184–186], cancer de la région tête/cou[187,178], cancer oral[188], de la prostate[189], de la peau[190], de la thyroïde[191,192], entres autres. Cette délétion est identique à celle observée dans le cadre du syndrome de Kearn-Sayre[171]. Cette délétion supprime sept gènes impliqués dans la chaîne respiratoire, les gènes *ATP6*, *ATP8*, *COX3*, *ND3*, *ND4L*, *ND4*, et *ND5*, ainsi que cinq gènes codant pour des ARNs de transfert mitochondriaux. Cette délétion impacte donc très fortement le fonctionnement de la phosphorylation oxydative.

En plus des mutations acquises au sein du génome mitochondrial dans les cellules tumorales, d'importants changements ont été observés au niveau du nombre de mitochondries par cellule tumorale, et ce dans de très nombreux cancers[142]. Dans les cancers de l'endomètre, de l'oesophage, colorectal, de la tête et du cou, de l'ovaire, carcinome papillaire de la thyroïde, et cancer de la prostate, le nombre de copies de mitochondries a tendance à être augmenté, alors que dans les cancers de l'estomac, du sein, carcinome fibrolamellaire, carcinome hépatocellulaire, sarcome d'Ewing, une diminution nette du nombre de mitochondries est observée, et ce de manière robuste.

Les altérations génomiques observées dans le cadre du développement tumoral sont la conséquence de l'inefficacité ou de l'absence de prise en charge des dommages occasionnés à l'ADN, que ce soit l'ADN nucléaire ou mitochondrial. Ces attaques peuvent avoir diverses origines, telles que les radiations ionisantes, mais la plus grande partie des attaques sont perpétrées par des composés appelés espèces oxygénées réactives, ou ROS - *Reactive Oxygen Species*.

IV.2 Espèces oxygénées réactives

IV.2 .1 Description

Les espèces oxygénées réactives, ou ROS, sont des composés dérivés du métabolisme de l'oxygène, et qui possèdent très souvent des électrons non appariés ce qui les rend extrêmement réactives. En effet, afin d'acquérir plus de stabilité électrochimique, elles réagissent avec de nombreux composés cellulaires, tels que les membranes lipidiques, les protéines cellulaires, et les acides nucléiques tels que l'ADN.

Une partie importante de l'oxygène que nous respirons conduit à la formation de molécules d'eau en subissant une réduction tétravalente soit par addition de quatre électrons (Équation 1). Cette réaction est catalysée par la cytochrome oxydase, accepteur terminal d'électrons présent dans le complexe IV de la chaîne respiratoire située dans la membrane interne des mitochondriale.

(1) $\mathrm{O_2 + 4\,e^- + 4\,H^+ \longrightarrow 2\,H_2O}$

Cependant, cette réduction de l'oxygène n'est pas parfaitement efficace, et une petite proportion d'électrons s'échappe de la chaîne de transport. Ainsi, environ 2% subit une réduction monoélectronique (ajout d'un seul électron, réaction 2), et conduit à la formation du radical superoxyde au niveau du coenzyme Q[193]. La NADH-déshydrogénase, située au niveau du complexe I de la chaîne respiratoire au niveau de la membrane interne mitochondriale peut également conduire à la formation du radical superoxyde, tout comme l'auto-oxydation (oxydation par l'oxygène) de coenzymes réduits tels que $FADH_2$ au niveau du complexe II.

(2) $\mathrm{O_2 + e^- \longrightarrow O_2^{\cdot -}}$

La disparition du radical superoxyde est effectuée par des enzymes appelées superoxyde dismutases ou SOD, qui catalysent sa dismutation (Équation 3) en une molécule de dioxygène et une molécule de peroxyde d'hydrogène, H_2O_2. Le peroxyde d'hydrogène n'est pas par définition un radical libre mais une molécule ayant tous ses électrons appariés.

(3) $\mathrm{O_2^{\cdot -} + O_2^{\cdot -} \xrightarrow{SOD,2\,H^+} H_2O_2 + O_2}$

Le peroxyde d'hydrogène est un intermédiaire réduit de l'oxygène relativement toxique. Cette toxicité vient la capacité du peroxyde d'hydrogène à générer le radical hydroxyle $\cdot$OH en présence de cations métalliques tels que Fe_2^+ ou Cu^+, réaction appelée réaction de Fenton :

(4) $\mathrm{H_2O_2 + Fe_2^+ \longrightarrow {\cdot}OH + Fe_3^+ + OH^-}$

Le radical hydroxyl $\cdot$OH et l'anion basique HO^- sont tous les deux formés lors de la réaction de Fenton. Cependant, alors que le second a tous ses électrons appariés, le premier possède un électron célibataire sur sa couche périphérique, ce qui le rend extrêmement plus réactif avec les composés environnants. Les radicaux hydroxyles sont les espèces oxygénées réactives les plus dommageables en raison de leur extrême réactivité. Ils attaquent tous les composés biologiques cellulaires : acides nucléiques, protéines, lipides. Ce sont des oxydants puissants qui réagissent

soit en arrachant un électron (réaction 5), soit en arrachant un atome d'hydrogène d'un substrat organique RH (réaction 6), soit en s'additionnant sur les doubles liaisons (réaction 7).

(5) $\cdot OH + Fe_2^+ \longrightarrow Fe_3^+ + {}^-OH$

(6) $\cdot OH + RH \longrightarrow R^{\cdot} + H_2O$

(7) $\cdot OH + \ > C = C < \ \longrightarrow \ \cdot > C{-}C(OH)^-$

Paradoxalement, les radicaux superoxydes sont eux-mêmes peu réactifs vis-à-vis de la majorité des biomolécules. Leur toxicité s'exerce de manière indirecte. En effet, le peroxyde d'hydrogène réagit avec le superoxyde pour générer des radicaux $\cdot OH$, réaction 8 appelée réaction Haber Weiss, catalysée par Fe_3^+ :

(8) $H_2O_2 + O_2^{\cdot -} \xrightarrow{Fe_3^+} \cdot OH + O_2 + OH^-$

IV.2 .2 Sources

D'où viennent et comment sont créés ces composés cytotoxiques ? Les sources d'espèces réactives oxygénées peuvent être classées en deux catégories : les sources endogènes à notre organisme, et les sources exogènes, provenant de l'environnement.

La mitochondrie est l'un des premiers contributeurs à la charge cellulaire en espèces oxygénées réactives, et ce de par l'activité des complexes I, II et III de la chaîne respiratoire permettant la production d'ATP.

La NAPDH oxydase est une enzyme catalysant la production de radicaux superoxyde à partir de l'oxygène et de NAPDH selon la réaction 9. Cette enzyme est active essentiellement au niveau des phagocytes, des cellules capables de prendre en charge et de dégrader des éléments pathogènes[194].

(9) $NAPDH + 2\,O_2 \longrightarrow NAPD^+ + H^+ + O_2^{\cdot -}$

Le réticulum endoplasmique est un organite cellulaire constitué d'un réseau de membranes interconnectées, dont la fonction est la production de protéines et de lipides. Le réticulum endoplasmique est également impliqué dans la détoxification cellulaire après exposition à des produits chimiques, et ce par l'intermédiaire d'enzymes mono-oxygénases de la famille cytochrome P450[195].

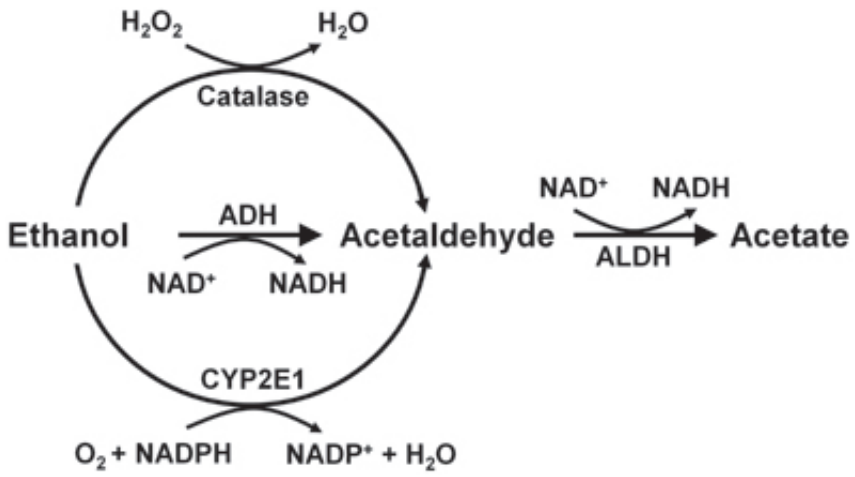
Les peroxysomes sont un type de compartiment cellulaire contenant des enzymes capables d'oxyder un grand nombre de molécules organiques, et sont fortement actifs au niveau du foie et des reins. Parmi ces enzymes, beaucoup génèrent des radicaux superoxydes ou du peroxyde d'hydrogène comme produit de leur activité catalytique normale. De ce fait, ils sont une importante source d'H_2O_2 cellulaire[196].

L'exposition à certains facteurs environnementaux organiques ou non organiques active les voies métaboliques décrites ci-dessus et par ce biais augmentent la quantité d'espèces oxygénées réactives générées. Ainsi, l'exposition à la pollution atmosphérique et en particulier aux

particules fines favorise la formation de radicaux libres[197,198]. Certains pesticides sont connus pour induire une cytotoxicité due aux espèces oxygénées réactives[199]. Des composés utilisés dans l'industrie tels que le chrome et le cadmium sont cytotoxiques[200]. C'est également le cas pour d'autres métaux tels que le mercure, le plomb, l'arsenic. Les composés appartenant à la famille des xenoestrogènes, appelés également perturbateurs endocriniens, augmente le taux de ROS cellulaires en inhibant l'activité des enzymes protectrices régulatrices du niveau d'espèces réactives oxygénées. Ces composés sont entre autres les phtalates, le paraben, ou le bisphénol A, et sont présents dans toute une variété de produits, tels que les plastiques, les emballages alimentaires, ou les produits cosmétiques et pharmaceutiques[201]. L'inhalation de la fumée de cigarette et le tabac en général sont également une source de ROS[202].

Par ailleurs, il a été montré que la consommation d'alcool augmente la production d'espèces réactives oxygénées. Lorsqu'on évoque l'alcool, on parle de l'espèce chimique appelée éthanol. Il existe plusieurs voies métaboliques de dégradation de l'éthanol au sein d'une cellule (voir Figure 25).

Figure 25 – Mécanismes de dégradation de l'éthanol



ADH : Alcool Déshydrogénase
ALDH : Acetaldéhydedéshydrogénase
D'après Israel *et al.*, Frontiers in Behavioral Neuroscience, 2013[203].

La première fait intervenir une enzyme appelée Alcool-Déshydrogénase (ou ADH), active dans une très grande majorité de nos cellules, qui dégrade l'éthanol en acetaldéhyde à partir de NAD^+. C'est la voie de dégradation par défaut, et la plus utilisée. Dans certains organes comme le cerveau, l'éthanol est dégradé par une enzyme appelée catalase, active dans les peroxysomes (voir section IV.3). Enfin, une consommation chronique et importante d'alcool provoque la dégradation d'une partie de l'éthanol par le cytochrome P450 CYP2E1, voie également active dans le cerveau. L'acetaldéhyde est à son tour dégradé en acétate par l'enzyme acetaldéhyde déshydrogénase dans la mitochondrie. De par leur activité, les enzymes alcool déshydrogénase

et acetaldéhyde déshydrogénase produisent toutes les deux du NADH, qui est lui même substrat de la NADH-déshydrogénase, le complexe I de la chaîne respiratoire. Ainsi, afin de conserver un ratio $\frac{NADH}{NAD+}$ à un niveau acceptable dans la cellule, NADH est réduit par la chaîne respiratoire, occasionnant au passage la synthèse de ROS.

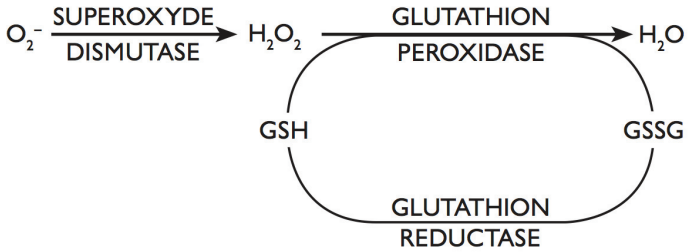
IV.3 Mécanisme de défense de l'organisme contre les ROS

Notre organisme possède des outils de défense face à ces espèces oxygénées réactives. Les principaux moyens de défense sont des enzymes capables de dégrader ces composés, mais il existe d'autres composés exogènes qui contribuent à limiter l'abondance et les dégâts occasionnés par les ROS.

Les superoxyde dismutases sont un ensemble d'enzymes capables de catalyser la dismutation du superoxyde en dioxygène et peroxyde d'hydrogène, comme vu précédemment à l'équation 3. Les superoxyde dismutases sont des métalloprotéines caractérisées par le métal contenu au niveau de leur site actif et nécessaire à leur activité catalytique. Chez l'Homme, il existe trois superoxyde dismutases[204]. La première identifiée, CuZnSOD, contient du cuivre et du zinc et est observée dans le cytoplasme cellulaire. La seconde et la plus récemment caractérisée contient également du cuivre et zinc, mais est localisée exclusivement dans des compartiments extracellulaires, elle est appelée ECSOD pour *Extra-Cellular SuperOxyde Dismutase*. Enfin MnSOD, la troisième, contient du manganèse, et est localisée dans la mitochondrie.

Les gluthation perxoydases (GPxs) sont une famille d'enzymes capables de catalyser la réduction de H_2O_2 en eau en utilisant la forme réduite du glutathion GSH comme réducteur :

Figure 26 – Cycle de dégradation de $O_2^{\cdot-}$ et de H_2O_2 par les enzymes antioxydantes superoxyde dismutase, glutathion peroxydase, et glutathion réductase.
D'après Pinto et al., Down Syndrome Research and Practice, 2002[205].



Il existe 8 formes de GPx[206] observées chez l'Homme, et certaines d'entre elles (GPx1, 2, 3, 4 et 6) contiennent du sélénium au niveau de leur site catalytique. La première forme découverte[207], GPx1, est active dans quasiment tous les tissus du corps humain, alors que les sept autres enzymes sont tissu-spécifiques. Par exemple, GPx2 est principalement trouvée au niveau de l'épithélium intestinal[208]. GPx5 est elle active au niveau de l'appareil reproducteur

masculin dans l'épididyme[209]. Lors de la transformation du peroxyde d'hydrogène en eau, les GPxs oxydent le glutathion en sa forme oxydée GSSG. Afin qu'une quantité suffisante de GSH soit disponible en permanence, une autre enzyme, la glutathion réductase catalyse la réaction inverse, soit la réduction de GSSG en GSH[210]. Ces trois familles d'enzymes, superoxyde dismutase, glutathion peroxydase, et glutathion réductase, constituent la principale voie de prise en charge des espèces oxygénées réactives que sont le radical superoxyde $O_2^{\bullet -}$ et le peroxyde d'hydrogène H_2O_2.

La catalase (évoquée précédemment dans la détoxification de l'éthanol) participe également à la protection de l'organisme contre le peroxyde d'hydrogène en catalysant sa dismutation en oxygène et en eau, selon la réaction 10. On trouve la catalase dans les peroxysomes. Le site actif de la catalase possède un cofacteur appelé hème, structure aromatique comprenant un atome de Fer couramment utilisée pour effectuer un transfert d'électrons.

(10) $2\,H_2O_2 \longrightarrow O_2 + 2\,H_2O$

Le terme d'antioxydant désigne également l'ensemble des composés naturels présents dans notre alimentation, favorisent l'action des enzymes antioxydantes ou neutralisant eux-mêmes certaines espèces oxygénées réactives[211,212]. Parmi les antioxydants non enzymatiques, plusieurs classes de composés ont été répertoriés parmi lesquels on peut citer les flavonoïdes, présents dans de très nombreux fruits et légumes mais également dans le chocolat, le vin rouge, et le thé. Ces composés sont très riches en polyphénols qui contribuent à la métabolisation des radicaux libres dans l'organisme. On trouve également parmi les antioxydants : les vitamines C, D et E dérivées des plantes, l'acide urique, certains minéraux essentiels aux bon fonctionnement des enzymes concernées comme le cuivre, le manganèse, le fer, le sélénium et le zinc. Enfin, en font également partie les caroténoïdes, ces pigments naturels apportant une coloration allant du vert à l'orangé rouge. Le principal caroténoïde connu est le bêta-carotène (carottes), mais la lutéine et le lycopène (respectivement dans les épinards et les tomates) sont également de puissants antioxydants.

IV.4 Mitochondrie, Stress Oxydatif et Cancer

Les espèces oxygénées réactives ont la capacité d'occasionner des dommages chimiques pouvant compromettre l'intégrité de nos cellules, et en particulier celle de notre ADN. Notre corps dispose de moyens de défense pour s'en protéger. Cependant, il arrive que ces moyens soient insuffisants pour neutraliser ces composés. La cellule est alors dans un état dans lequel ses constituants et notamment les acides nucléiques comme l'ADN, subissent des attaques chimiques plus ou moins sévères, cet état est appelé stress oxydatif. Les types de lésions occasionnées à l'ADN peuvent être entre autres, l'oxydation des bases nucléiques, la création de site abasiques - sites au niveau desquels une base nucléique est manquante - ou encore des cassures simple-brin[213]. Plus de 20 types de lésions pouvant être occasionnées sur les bases de l'ADN ont été répertoriés[214,215]. La base nucléique guanine est la plus facilement oxydable.

Pendant longtemps la mitochondrie n'a été que la centrale énergétique de la cellule, mais aujourd'hui il est évident que cet organite dynamique joue un rôle dans la carcinogenèse. En effet, les altérations génétiques et métaboliques au niveau de la mitochondrie ont été montrées comme étant la cause ou un facteur contributeur d'un certain nombre de pathologies humaines, incluant le cancer[216,217]. Le stress oxydatif est le biais par lequel le génome mitochondrial pourrait contribuer à influer sur le risque de cancer. Comme vu précédemment, une des sources majeures de ROS dans la cellule est la phosphorylation oxydative. Les espèces oxygénées réactives générées par la mitochondrie peuvent favoriser le développement du cancer par deux biais : en premier lieu, ils peuvent causer des mutations dans des gènes nucléaires d'importance capitale (proto-oncogènes, gènes suppresseurs de tumeurs, gènes de réparation de l'ADN), et d'autre part, en agissant comme un facteur promoteur du développement tumoral en favorisant la prolifération cellulaire. En effet, à des taux raisonnables (suffisamment faibles pour ne pas conduire la cellule à l'apoptose), les ROS interagissent avec une quantité de protéines (notamment les protéines kinases) et de facteurs de transcription, conduisant à la division cellulaire[218].

Le génome mitochondrial est lui-même d'autant plus à même de subir des dommages dus au stress oxydatif que celui-ci est directement exposé aux espèces oxygénées réactives générées par la phosphorylation oxydative au niveau de la membrane mitochondriale interne. Dans l'hypothèse ou une tumeur commencerait à se développer et où ses mécanismes de contrôle de l'intégrité et réparation de l'ADN seraient altérés, il n'est pas surprenant d'observer une instabilité génomique dans la tumeur au niveau de l'ADN mitochondrial comme au niveau de l'ADN nucléaire.

Des mutations constitutionnelles dans des gènes portés par l'ADN mitochondrial ont été identifiées comme prédisposant au cancer. Certaines mutations affectant le gène mitochondrial *COI*, et plus précisément les mutations T6253C, C6340T, G6261A, et A6663G augmentent fortement le risque de cancer de la prostate[216]. Dans une étude analysant le génome mitochondrial d'individus mâles dont 180 atteints d'un cancer de la prostate, 46 contrôles négatifs âgés de plus de 50 ans, et de 898 génomes mitochondriaux représentant la population générale, ces mutations du gène *COI* ont été observées dans 11% des cas de cancer de la prostate, 0% des contrôles, et 5.5% de la population générale, avec une fréquence statistiquement différente entre chacun de ces groupes[216]. Plusieurs études ont d'autre part mis en évidence le rôle primordial des mutations constitutionnelles et somatiques dans le développement et la progression du cancer de la prostate[216,219]. Des cellules d'une lignée de cancer de la prostate ont été déplétées de leurs mitochondries, puis séparées en deux groupes. On a ensuite injecté au premier groupe des mitochondries possédant la mutation T8993G dans le gène *ATP6*, alors que des mitochondries sans cette mutation ont été injectées au second groupe. Cette mutation avait précédemment été étudiée pour son rôle potentiel dans la surproduction de ROS[220]. Ces cellules, possédant le même génome nucléaire, mais un génome mitochondrial variable fusionné de manière artificielle, sont appelées cybrides, terme provenant de la contraction des mots cytoplasme et hybride. Ces hybrides ont ensuite été injectés à des souris. Les souris ayant reçu les cybrides sans la mutation T8993G ont eu une croissance normale. Les souris ayant reçu les cybrides mutés ont développé rapidement des tumeurs conduisant à leur mort. De plus, il a été observé que les cellules arborant cette mutation produisent beaucoup plus de ROS que les cellules ne la présentant pas[220,216].

Le cancer du sein semble lui aussi être caractérisé par des dysfonctionnements des fonctions mitochondriales dus à des altérations génétiques de la phosphorylation oxydative. Les protéines structurelles impliquées dans la chaîne de transport des électrons sont encodées aussi bien par des gènes nucléaires que par des gènes portés par le génome mitochondrial. Certaines mutations affectant la séquence de la D-loop pourraient également favoriser le développement du cancer du sein.

L'allèle A du polymorphisme situé en 10398 a été montré augmentant le risque de cancer du sein dans certaines populations : chez les femmes afro-américaines mais pas chez les femmes caucasiennes[221], et chez les femmes originaires du nord de l'Inde[222]. Cependant, d'autres études n'ont pas détecté cette association[223,224] dans ces populations. Cet allèle a également été observé associé avec le risque de cancer du sein dans des populations européennes, américaines d'origine caucasienne, polonaises, et malaysiennes[225–228]. Cette association est à l'heure actuelle toujours controversée. De plus, une étude a envisagé une éventuelle modification de l'effet de l'association de l'allèle 10398A par la consommation d'alcool[229]. Ce polymorphisme non-synonyme pourrait altérer l'efficacité et le fonctionnement du complexe I de la chaîne respiratoire, et ainsi augmenter significativement la quantité d'espèces oxygénées réactives générées[230], favorisant ainsi le développement tumoral. La consommation d'alcool accentuerait à un niveau supplémentaire l'excès de ROS générés.

Le polymorphisme T16189C a été détecté associé avec le risque de cancer du sein[231], tout comme G9055A, T16519C, T239C, A263G, and C16207T[225,232]. Les polymorphismes T3197C et G13708A ont quant à eux été détectés inversement associés avec le risque de cancer du sein[225], alors que les SNPs A73G, C150T, T16183C, T16189C, C16223T, et T16362C ont été observés avec une fréquence plus élevée chez les cas de cancer du sein que chez des contrôles sains[232]. Récemment, 69 nouveaux variants ont également été découverts, et l'haplogroupe M5 a été observé associé avec le risque de cancer du sein[233].

Ainsi, la variabilité du génome mitochondrial, en particulier au niveau des gènes codant pour des sous-unités structurelles de la chaîne respiratoire, semblerait jouer un rôle dans la prédisposition génétique au cancer, notamment au cancer du sein, en influant sur la quantité d'espèces oxygénées réactives produites.

L'ensemble des éléments présentés dans cette introduction illustre l'existence de données de nature épidémiologique et biologique selon lesquelles le génome mitochondrial pourrait avoir une influence sur le risque de cancer, et notamment de cancer du sein. Il existe des syndromes pathologiques graves causés par des mutations au sein du génome mitochondrial, mutations ayant un effet dramatique sur le fonctionnement de la mitochondrie et donc de la cellule. Dans ce contexte, il parait donc légitime de s'interroger sur la possibilité que d'autres mutations localisées sur le génome mitochondrial puissent contribuer ou influer sur le risque de cancer. D'autre part, en comparaison du génome nucléaire, le génome mitochondrial n'a été que peu étudié, et ce principalement parce qu'il présente des spécificités qui font que les technologies employées pour l'analyse du génome nucléaire ne sont pas optimisées pour son étude. De ce fait, bien que des études sur les liens entre mitochondrie et cancer aient été publiées, la littérature sur le sujet est relativement peu fournie. Dans ce contexte, les travaux que j'ai effectués au

cours de mon doctorat ont pour objectif de contribuer à pallier à ce manque de données et d'approfondir l'étude des liens existants entre génome mitochondrial et cancer du sein.

V. Présentation des travaux de thèse

La suite de ce manuscrit présente les travaux que j'ai réalisés. Pendant mon doctorat, je me suis intéressée à la variabilité du génome mitochondrial en tant que facteur de prédisposition génétique au cancer du sein. J'ai abordé cette question selon trois axes de recherche. Chaque étude, les données et les méthodes utilisées, ainsi que les résultats obtenus seront successivement présentés. Ils seront commentés de manière critique et replacés dans leur contexte dans une discussion à l'issue de chaque analyse. Enfin, une discussion générale faisant le lien entre les différents axes de recherche mis en oeuvre et les résultats obtenus sera présentée.

Le premier axe m'a conduit à étudier une association entre le risque de cancer du sein et une interaction potentielle entre des variants du génome nucléaire localisés dans la séquence codante d'enzymes protectrices contre les espèces réactives oxygénées, et entre un variant du génome mitochondrial et l'exposition au facteur environnemental relatif au style de vie qu'est la consommation d'alcool. Ces analyses ont été réalisées sur un jeu de données issu du consortium international *Breast and Prostate Cancer Cohort Consortium*, ou BPC3.

Le second axe de recherche que j'ai développé porte sur l'identification de sous-haplogroupes mitochondriaux ayant un risque modifié de cancer du sein par rapport aux clades avoisinantes ainsi qu'en comparaison de la population générale, chez des femmes présentant un historique familial de cancer du sein et portant une mutation pathogène sur un des deux principaux gènes de prédisposition au cancer du sein à forte pénétrance, *BRCA1* et *BRCA2*. Une approche innovante basée sur l'évolution du génome mitochondrial a, de plus, été appliquée afin d'augmenter la puissance statistique des analyses. Les données analysées dans le cadre de cette étude proviennent du projet COGS, pour *Collaborative Oncological Gene-environment Study*.

Le troisième axe de recherche développé a pour but de caractériser la variabilité du génome mitochondrial chez des femmes diagnostiquées pour un cancer du sein et ayant un historique familial pour cette maladie, mais testées négatives pour les mutations pathogènes sur *BRCA1* et *BRCA2*. Le génome mitochondrial de 436 femmes appartenant à l'étude GENESIS a été séquencé en utilisant la technologie Ion Torrent, et l'analyse bioinformatique de ces données est présentée.

Chapitre 1

Étude de facteurs associés au stress oxidatif et risque de cancer dans le cadre du Breast and Prostate Cancer Cohort Consortium

Comme il a été évoqué précédemment, la manganèse superoxide dismutase MnSOD et la glutathion peroxidase GPx1 sont deux enzymes appartenant à la catégorie des antioxydants et qui, par leur action sur la régulation des espèces oxygénées réactives, protègent les cellules et l'ADN du stress oxydatif. Des variations dans la séquence des gènes associés *MnSOD* et *GPx1* peuvent modifier la composition en acides aminés des protéines qu'ils encodent, et les fonctions et l'efficacité de régulation de MnSOD et GPx1 peuvent en être altérées[234–236]. Une étude réalisée au sein de la cohorte *Nurses' Health Study*[237] - ou NHS - a montré que les individus porteurs homozygotes d'une Alanine en position 16 de la protéine MnSOD (Ala16Ala, polymorphisme rs4880) et porteurs homozygotes d'une Leucine en position 198 de la protéine GPx1 (Leu198Leu, polymorphisme rs1050450) ont un risque de cancer du sein multiplié par 1.87 comparé aux porteurs d'au moins une Valine et d'une Proline aux positions correspondantes respectives sur MnSOD et GPx1. Les deux polymorphismes étudiés auraient donc un impact synergique sur le risque de cancer du sein.

Une consommation excessive d'alcool est d'autre part un facteur lié au style de vie qui augmente le risque de cancer du sein. Une étude[229] a cependant observé que l'effet de l'alcool sur le risque de cancer du sein était modifié par le polymorphisme A10398G situé sur le gène *ND3*, localisé sur le génome mitochondrial. Ce gène code pour la troisième sous-unité du complexe I de la chaîne respiratoire mitochondriale. Ce polymorphisme a lui aussi été génotypé dans l'étude *Nurses' Health Study*, et également dans l'étude *Women's Health study*. L'analyse des données a conclu que, alors que l'alcool n'avait pas d'effet sur le risque de cancer du sein chez les porteurs de l'allèle A du polymorphisme situé au locus 10398 du génome mitochondrial, un effet était observé chez les porteurs de l'autre allèle. Ainsi, d'après cette étude, les porteurs de l'allèle G consommateurs d'alcool ont un risque de cancer du sein multiplié par 1.52 (Intervalle de Confiance à 95% 1.10-2.08) en comparaison des individus ne consommant pas d'alcool.

En statistiques, lorsqu'on réalise un test, la décision de rejeter l'hypothèse nulle H_0 est prise en contrôlant le risque de 1$^{\text{ère}}$ espèce α que l'on a de prendre cette décision à tort. Mais même si on sait précisément quel risque on prend, et que ce risque est généralement faible, il est possible que la distribution des données ayant conduit au rejet de H_0 soit en réalité observée uniquement par le fruit du hasard. Les études d'association effectuées dans le domaine de l'épidémiologie génétique ne se basent que sur des différences de distribution statistique, et ne s'appuient sur aucune preuve fonctionnelle. En effet, c'est souvent après avoir détecté une association génétique que les mécanismes sous-jacents sont élucidés et que les altérations caractéristiques de la pathologie étudiée sont identifiées. Ainsi, les études épidémiologiques publiées mettant en

évidence une association pour la première fois ne contiennent généralement pas d'éléments de validation fonctionnelle étayant les conclusions établies. C'est pourquoi les conclusions d'une étude ne sont généralement considérées comme robustes que lorsqu'elles ont été répliquées avec succès. La réplication d'une étude consiste à mettre en place une seconde étude ayant le même design que l'étude initiale, mais effectuée sur une population indépendante, et présentant des caractéristiques comparables (ethnie, exposition, âge, état de santé...) . Si l'association détectée initialement est réelle, et que la seconde étude a la puissance statistique suffisante, alors l'association devrait également être détectée à l'issue de la phase de réplication.

Le travail réalisé dans cette première partie a pour but de répliquer les résultats des deux études présentées[229,237], afin de valider d'une part l'interaction entre les deux polymorphismes non-synonymes contenus dans les enzymes MnSOD et GPx1, et d'autre part la modification de l'effet de l'alcool en fonction du génotype du polymorphisme mitochondrial contenu situé sur le gène *ND3*. Cette réplication a été effectuée dans le cadre du *Breast and prostate Cancer Cohort Consortium*, ou BPC3. Comme son nom l'indique, le BPC3 a pour objectif d'étudier l'implication de facteurs génétiques et de facteurs liés au style de vie dans la prédisposition aux cancers du sein et de la prostate. Les analyses répliquées dans le cadre du cancer du sein ont donc été mises en places de la même manière dans le cadre du cancer de la prostate, afin de tester si les associations précédemment détectées étaient également observées dans cet autre type de cancer. Des analyses de survie ont également été mises en place afin de déterminer l'influence des facteurs étudiés sur la mortalité des individus après leur diagnostic.

I. Matériels et Méthodes

I.1 Le *Breast and Prostate Cancer Cohort Consortium*

Le *Breast and Prostate Cancer Cohort Consortium*, ou BPC3 est un consortium international financé et piloté par l'institut national de santé américain dédié au cancer, le *National Cancer Institute*, ou NCI. Ce consortium mutualise les ressources de neuf études de cohortes préexistantes : l'étude *Alpha-Tocopherol, Beta-Carotene Cancer Prevention* (ATBC), l'étude *American Cancer Society Cancer Prevention Study II* (CPS-II) , *European Prospective Investigation into Cancer and Nutrition Cohort* (EPIC) (elle-même composée d'études originaires de France, Danemark, Royaume-Uni, Allemagne, Grèce, Italy, Pays-Bas, Espagne et Suède), l'étude *Health Professionals Follow-up Study* (HPFS), l'étude *Multiethnic Cohort* (MEC), l'étude *Physicians' Health Study* (PHS), l'étude *Nurses' Health Study* (NHS), l'étude *Women's Health Study* (WHS), et l'étude *Prostate, Lung, Colorectal, and Ovarian Cancer screening trial* (PLCO). Chacune de ces études a été approuvée par les comités d'examens locaux responsables, et tous les participants ont fourni leur consentement éclairé.

Chacune de ces études possède des critères qui lui sont propres que ce soit pour le recrutement des cas, la mesure de l'exposition, ou encore l'appariement des contrôles[238]. Au total, 13 511 femmes atteintes d'un cancer du sein et 8 490 hommes atteints d'un cancer de la prostate ainsi que des contrôles ont été inclus dans une première analyse visant à étudier l'interaction entre deux polymorphismes localisés respectivement dans les gènes *MnSOD* et *GPx1*. De même, 10 726 femmes atteintes d'un cancer du sein ainsi que 7 532 hommes atteint d'un cancer de la

prostate ainsi que des contrôles ont été inclus dans une seconde analyse visant à étudier l'interaction potentielle entre le polymorphisme mitochondrial rs2853826/A10398G et la consommation d'alcool.

Ces deux populations d'études sont globalement représentatives de la population générale en ce qui concerne le statut tumoral des récepteurs hormonaux. Dans le cas du cancer du sein, comme l'ont décrit Hendrickson et ses collaborateurs[239], le statut tumoral de sensibilité aux oestrogènes et à la progestérone (statut ER/PR) est connu pour environ 60% à 80% des participants. Le statut ER observé suit la distribution générale, avec environ 20% de femmes dont le cancer est ER négatif - les cellules tumorales ne possèdent pas de récepteurs aux oestrogènes à leur surface - et 80% de femmes dont le cancer est ER positif - les cellules tumorales possèdent des récepteurs aux oestrogènes à leur surface. Le statut des participants concernant des biomarqueurs tumoraux tels que le taux d'expression du gène EGFR ou l'amplification de l'expression gène HER2 n'est pas connu. En effet, une grande partie des cas de cancer du sein ont été recrutés avant que ces biomarqueurs ne soient introduits dans la batterie d'analyses réalisés en routine au sein des cohortes analysées.

I.2 Génotypage

Trois SNPs ont été génotypés : rs1050450 et rs4880, situés respectivement sur les gènes nucléaires *MnSOD* et *GPX-1*, et rs2853826 situé sur le gène mitochondrial *ND3*. Pour la cohorte PLCO, le SNP rs8031 sur le gène *MnSOD* a été génotypé à la place du SNP rs4880. Ces deux polymorphismes étant en très fort déséquilibre de liaison ($r^2 = 0.95$ d'après les fréquences génotypiques issues du projet HapMap), le génotype de rs8031 a été utilisé pour inférer celui de rs4880.

Le génotypage a été effectué par TaqMan. Les SNPs rs4880 et rs1050450 ne dévient pas de l'équilibre de Hardy-Weinberg chez les contrôles au sein de chaque étude. Au total, 334 individus ont été exclus des analyses à cause de la mauvaise qualité de leurs génotypes. Les cohortes pour lesquelles le taux de succès de génotypage d'un SNP donné n'atteignait pas les 90% ont été exclues des analyses pour le SNP considéré. Ainsi, pour chaque SNP et chaque cohorte, le taux de génotypage était supérieur à 0.91.

I.3 Analyses statistiques

Les analyses ont été effectuées sous R. Une régression logistique a été effectuée pour chaque type de cancer et pour chaque interaction testée. On teste donc si une interaction existe entre les facteurs de risques inclus dans chaque modèle, c'est à dire si le risque en présence des deux facteurs de risque est supérieur au produit des risques de chacun des facteurs individuellement. Sous R, le modèle générique correspondant est le suivant :

```
model = glm(response ~ factor1*factor2 + {adjustment_factors})
```

Dans le cas de l'interaction entre les SNPs situés sur les gènes *MnSOD* et *GPX-1*, toutes les analyses ont été conduites sous l'hypothèse d'un modèle génétique récessif comme c'était le cas pour les études précédentes[237,229]. Les facteurs d'ajustement pris en compte pour chaque

analyse sont présentés dans la Table 3. Les tests de Wald et du rapport de vraisemblance ont été effectués afin d'établir la significativité statistique des interactions. Les calculs de puissance statistiques ont été effectués à l'aide du Logiciel Quanto[240](Table 4).

Les analyses de survie sur les cas en fonction du génotype et des conditions de vie ont été effectuées en construisant des modèles de Cox à taux proportionnels avec le package R *Survival*[241]. Des tests d'hétérogénéité ont été effectués afin d'évaluer la similarité des associations entre toutes les cohortes inclues, et le modèle à effets mixtes a été retenu lorsque le test d'hétérogénéité était significatif ($p < 0.05$).

Une méta-analyse combinant nos résultats ainsi que les résultats publiés de la littérature concernant l'effet du polymorphisme rs1050450 (*GPx1*) sur le risque de cancer de la prostate a été mise en place. Six études ont été initialement sélectionnées[242–247]. Cependant, Steinbrecher et ses collaborateurs[247] ont basé leurs travaux sur des données provenant de la cohorte EPIC, elle-même appartenant au BPC3 sur lequel nos propres travaux se basent. À moins d'invalider l'hypothèse d'indépendance des données, nécessaire à la construction d'une méta-analyse correcte, cette étude a du être exclue de la méta-analyse. Les résultats de l'étude conduite par Cheng et ses collaborateurs furent également exclus car les informations relatives au statut homozygote ou hétérozygote des allèles portés n'étaient pas présentées. Les tests de Dixon et de Grubbs du package R *outliers* ont été utilisés afin de détecter la présence d'*outliers* parmi les études inclues dans la méta-analyse.

II. Résultats

II.1 Cancer du sein

L'interaction testée entre les polymorphismes rs4880 Val16Ala situé dans le gène *MnSOD* et rs1050450 Pro198Leu situé dans le gène *GPx1* est statistiquement non significative ($p = 0.34$) et ce polymorphisme ne semble pas avoir d'influence sur le risque de cancer du sein (Table 3). Cette étude a une puissance statistique supérieure à 95% pour détecter un odds ratio similaire à celui estimé par l'étude initiale. Cet odds ratio avait une valeur de 1.87, pour un risque de première espèce $\alpha = 0.05$, et dans le cas d'un modèle récessif impliquant deux polymorphismes n'ayant aucun effet seuls (Table 4). Le risque de cancer du sein chez les porteurs homozygotes de l'allèle alternatif pour les deux polymorphismes étudiés n'est statistiquement pas différent du risque des individus appartenant à la catégorie de référence (odds ratio = 1.03, pour un intervalle de confiance à 95% de [0.97 - 1.09]).

De même, le polymorphisme rs2853826 A10398G ne semble pas modifier l'association entre le risque de cancer du sein et la consommation d'alcool ($p = 0.98$). Les 2 études initiales répliquées ici étant effectuées sur les données provenant des cohortes NHS et WHS, nous avons vérifié avec succès que les résultats obtenus ne variaient pas en retirant ces deux cohortes de nos analyses.

Table 3 – Résultats des modèles d'association testés

Interaction	Catégorie	Cas	%	Témoins	%	OR	IC à 95%
Pro198Leu (*GPX1*) Val16Ala (*MnSOD*) Cancer du sein[a]	Pro198 - Val16	3215	66.3	3685	65.9	1	Ref.
	Pro198 - Ala16Ala	1134	23.4	1326	23.7	1.00	0.96 - 1.03
	Leu198Leu - Val16	371	7.6	442	7.9	1.00	0.97 - 1.02
	Leu198Leu - Ala16Ala	132	2.7	139	2.5	1.03	0.97 - 1.09
A10398G (*MT-ND3*) Alcool Cancer du sein[b]	A10398 - Pas d'alcool	1114	33.7	1443	36.2	1	Ref.
	A10398 - Alcool	1507	45.6	1732	43.5	1.13	1.02-1.26
	G10398 - Pas d'alcool	294	8.9	372	9.3	1.03	0.87-1.23
	G10398 - Alcool	391	11.8	436	10.9	1.16	0.99-1.36
Pro198Leu (*GPX1*) Cancer de la prostate[c]	Pro198	6688	89.4	6510	88.2	1	Ref.
	Leu198Leu	792	10.6	867	11.8	0.87	0.79-0.97
Pro198Leu (*GPX1*) Val16Ala (*MnSOD*) Cancer de la prostate[d]	Pro198 - Val16	4223	66.2	4230	66.3	1	Ref.
	Pro198 - Ala16Ala	1473	23.1	1396	21.9	1.06	0.97 - 1.15
	Leu198Leu - Val16	507	7.9	573	89	0.88	0.76 - 0.98
	Leu198Leu - Ala16Ala	176	2.8	208	3.3	0.84	0.67 - 1.02
A10398G (*MT-ND3*) Alcool Cancer de la prostate[e]	A10398 - Pas d'alcool	389	10.6	429	11.1	1	Ref.
	A10398 - Alcool	2448	66.7	2596	67.2	1.15	0.99 - 1.33
	G10398 - Pas d'alcool	135	3.7	131	3.4	1.12	0.85 - 1.48
	G10398 - Alcool	698	19.0	706	18.3	1.16	0.97 - 1.38

[a] P-value d'interaction : 0.34. Données restreintes aux femmes post-ménopause. Régression logistique non conditionnelle contrôlée sur l'âge au prélèvement sanguin, l'âge des premières menstruations, l'âge à la ménopause, l'indice de masse corporelle, l'histoire familiale de cancer du sein, et la cohorte.

[b] P-value d'interaction : 0.98. Données restreintes aux femmes post-ménopause. Régression logistique non conditionnelle contrôlée sur l'âge au prélèvement sanguin, l'âge des premières menstruations, l'âge à la ménopause, l'indice de masse corporelle, l'histoire familiale de cancer du sein, et la cohorte d'origine.

[c] Régression logistique non conditionnelle contrôlée sur l'âge au diagnostic, la consommation d'alcool, et la cohorte d'origine. P-value : 0.01

[d] P-value d'interaction : 0.44. Régression logistique non conditionnelle contrôlée sur l'âge au diagnostic, la consommation d'alcool, et la cohorte d'origine.

[e] P-value d'interaction : 0.50. Régression logistique non conditionnelle contrôlée sur l'âge au diagnostic, et la cohorte d'origine.

Table 4 – Calculs de la puissance statistique de réplication dans chacune des études mises en place

Interaction testée	Type de Cancer	Type d'analyse	Remarques	Risque absolu	Taille de l'échantillon	Ratio Témoins/Cas	Puissance
rs4880 × rs1050450	Sein	Gène×Gène	-	0.1	4852	1.152	Pour un OR = 1.87* : **P = 0.995**
	Prostate	Gène × Gène	-	0.1	6379	1.004	Pour un OR = 1.87* : **P = 0.999**
Alcool × rs2853826	Sein	Facteur environnemental	Individus G10398 uniquement	0.1	685	1.179	Pour un OR = 1.52** : **P = 0.976**
	Prostate	Facteur environnemental	Individus G10398 uniquement	0.1	808	1.045	Pour un OR = 1.52** : **P = 0.83**

* OR estimé dans l'étude initiale testant l'interaction entre rs4880 et rs1050450.
** OR estimé dans l'étude initiale testant l'interaction entre rs2853826 et la consommation d'alcool.

Les résultats des analyses de survie sont présentés dans la Table 5. Aucune différence de survie n'est observée entre les différentes catégories de génotypes des polymorphismes rs4880 et rs1050450. Cependant, une différence de survie totale est observée entre les groupes d'individus d'une part relativement à leur consommation d'alcool (Figure 27a), et d'autre part, en fonction de leur consommation d'alcool et de leur génotype au polymorphisme rs2853826 (Figure 27b). On ne retrouve pas cette différence de survie lorsqu'on considère la survie spécifique au cancer du sein, c'est à dire lorsqu'on étudie spécifiquement les décès dus au cancer du sein (Figure 27c). Par exemple, la survie spécifique au cancer du sein prend en compte les décès à la suite du développement métastatique du cancer initial, mais n'inclue pas les décès dus aux accidents de la route.

Table 5 – Résultats des analyses de survie

			Survie totale		Survie spécifique	
Cancer	Analyse	Catégorie	HR (IC95%)	LRT p-value	HR (IC95%)	LRT p-value
Sein	rs4880 × rs1050450 *	Pro198 & Val16	1 (Ref.)	0.747	1 (Ref.)	0.508
		Pro198 & Ala16Ala	0.93 (0.79 - 1.10)		1.09 (0.87 - 1.36)	
		Leu198Leu & Val16	0.93 (0.72 - 1.21)		0.85 (0.57 - 1.26)	
		Leu198Leu & Ala16Ala	0.85 (0.55 - 1.32)		0.74 (0.38 - 1.43)	
	Alcool × rs2853826*	A10398 & Pas d'Alcool	1 (Ref.)	0.029	1 (Ref.)	0.103
		A10398 & Alcool	0.9 (0.76 - 1.07)		1.24 (0.97 - 1.57)	
		G10398 & Pas d'Alcool	1.03 (0.77 - 1.36)		0.95 (0.62 - 1.44)	
		G10398 & Alcool	0.66 (0.49 - 0.88)		0.84 (0.57 - 1.26)	
	Alcool uniquement*	Pas d'Alcool	1 (Ref.)	0.003	1 (Ref.)	0.104
		Alcool	0.84 (0.74 - 0.94)		1.15 (0.97 - 1.36)	
Prostate	rs4880 × rs1050450**	Pro198 & Val16	1 (Ref.)	0.855	1 (Ref.)	0.508
		Pro198 & Ala16Ala	1.00 (0.89 - 1.12)		0.93 (0.77 - 1.14)	
		Leu198Leu & Val16	1.00 (0.84 - 1.18)		0.75 (0.54 - 1.04)	
		Leu198Leu & Ala16Ala	1.13 (0.86 - 1.49)		1.05 (0.64 - 1.70)	
	Alcool × rs2853826**	A10398 & Pas d'Alcool	1 (Ref.)	0.558	1 (Ref.)	0.148
		A10398 & Alcool	1.07 (0.87 - 1.32)		1.17 (0.81 - 1.67)	
		G10398 & Pas d'Alcool	0.85 (0.57 - 1.27)		0.49 (0.21 - 1.16)	
		G10398 & Alcool	1.09 (0.86 - 0.38)		1.13 (0.75 - 1.69)	

*Ajustement sur la cohorte d'appartenance et sur l'âge au diagnostique de cancer du sein

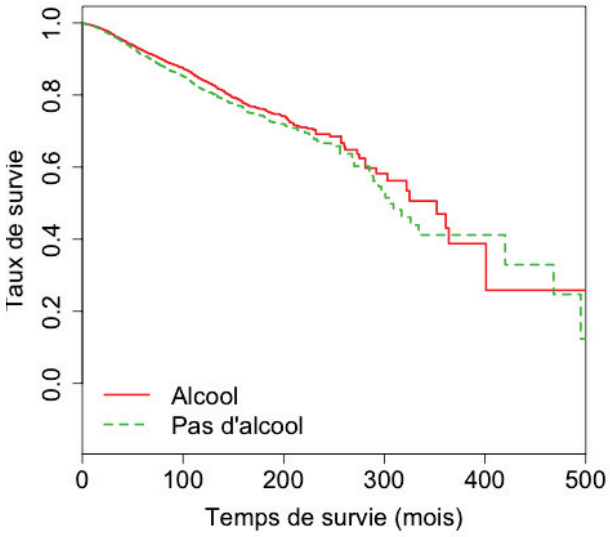
**Ajustement sur la cohorte d'appartenance et sur l'âge au diagnostique de cancer de la prostate

HR : hazard ratio

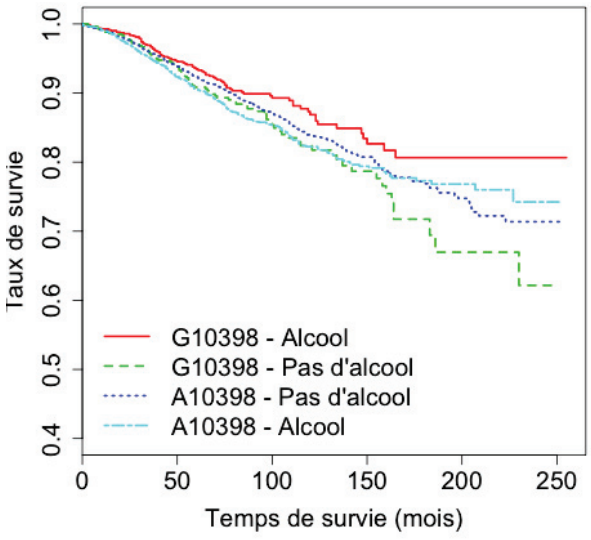
IC95% : Intervalle de confiance à 95%

LRT : LogRank test

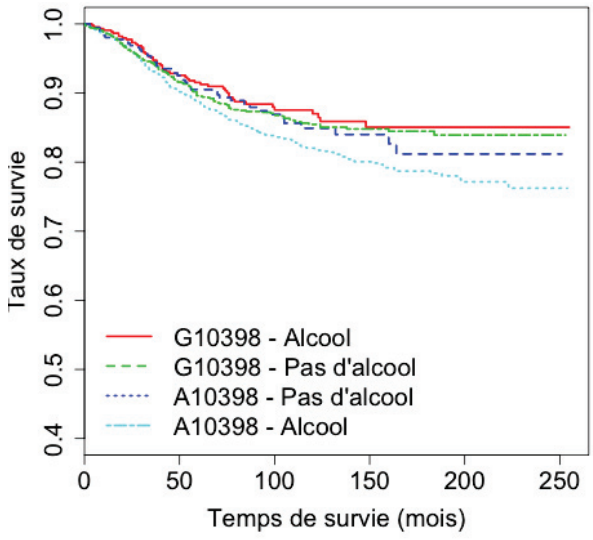
Figure 27 – Graphes de survie totale et spécifique du cancer du sein en fonction de la consommation d'alcool et du SNP A10398G



(a) Courbe de survie totale en fonction de la consommation d'alcool



(b) Courbe de survie totale en fonction de la consommation d'alcool et du SNP A10398G



(c) Courbe de survie spécifique du cancer du sein en fonction de la consommation d'alcool et du SNP A10398G

II.2 Cancer de la prostate

Une association inverse entre le risque de cancer de la prostate et l'allèle alternatif du SNP rs1050450 localisé dans le gène *GPx1* a été observée (Table 3). D'après les analyses effectuées, les porteurs homozygotes d'une Leucine en position 198 de la protéine GPx1 ont un risque de cancer de la prostate inférieur aux individus d'une autre classe de génotype, avec un odds ratio $OR = 0.87$, pour un intervalle de confiance à 95% de (0.79 - 0.97). L'interaction testée entre ce polymorphisme et rs4880 situé dans le gène *MnSOD* n'influe pas significativement sur le risque de cancer de la prostate. De même, aucune interaction n'a été détectée entre la consommation d'alcool et le polymorphisme rs2853826 du gène mitochondrial *ND3* dans le cadre de l'étude du risque de cancer de la prostate.

D'autres études ont déjà analysé l'interaction entre rs1050450 et le risque de cancer de la prostate[242,244–246]. Une méta-analyse a donc été effectuée afin de comparer nos résultats avec l'ensemble de ces études. Les résultats de cette méta-analyse sont présentés dans la figure 28. De l'hétérogénéité a été observée entre les études inclues, avec $I^2 = 74\%$, et $Q = 15.8$; la probabilité que $Q \hookrightarrow X^2_4$ est p=0.0033. Un modèle à effets mixtes a donc été sélectionné pour calculer l'odds ratio global, estimé à OR=1.19, (intervalle de confiance à 95% : 0.79 - 1.80). Cependant, l'application des tests de Dixon et Grubs ont permis d'identifier l'étude de Kucukgergin *et al.*[246] comme un outlier. Après exclusion de cette étude, une seconde méta-analyse a été effectuée et ne montre cette fois aucune hétérogénéité entre les études restantes. Le modèle à effets fixes mis en place dans cette seconde méta-analyse a permis d'estimer l'odds ratio global OR=0.90, (intervalle de confiance à 95% : 0.81 - 0.99).

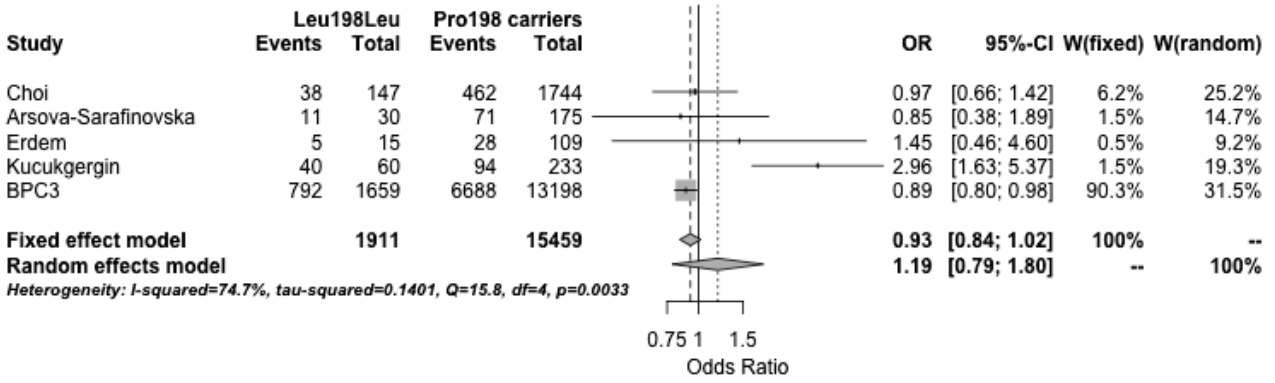
III. Discussion

L'objectif de cette analyse était d'approfondir l'étude de deux associations génomiques précédemment détectées[237,229] impliquant des polymorphismes potentiellement liés à un excès en espèces réactives oxygénées pouvant conduire à l'état de stress oxydatif dans le cadre des cancers du sein et de la prostate. Je me suis tout d'abord intéressée à des variations observés dans les gènes *MnSOD* et *GPx1*, 2 gènes encodant des enzymes antioxydantes qui protègent le contenu cellulaire et notamment l'ADN des dommages oxydatifs. Contrairement à l'étude initiale, les résultats obtenus montrent que les porteurs homozygotes des allèles alternatifs des polymorphismes r4880 Val16Ala dans *MnSOD* et rs1050450 Pro198Leu dans *GPx1* n'ont pas un risque de cancer du sein ou de la prostate modifié par rapport aux individus appartenant à une autre catégorie de génotype.

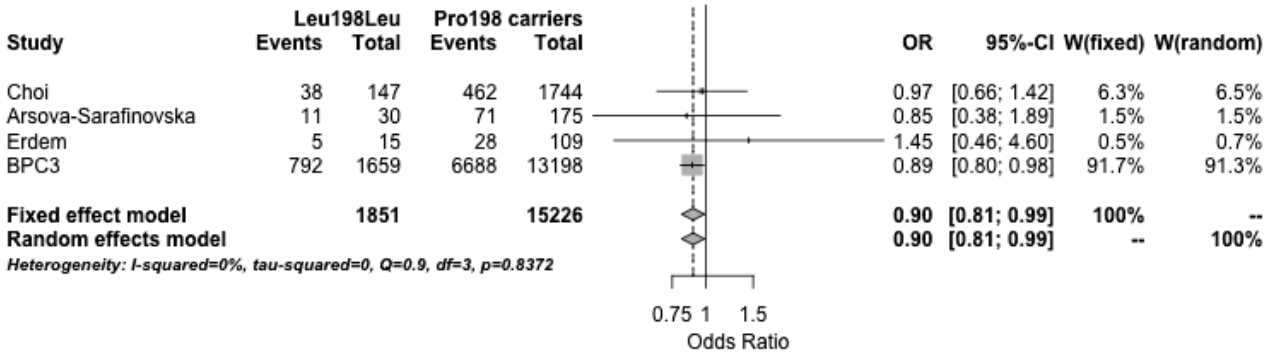
Nous avons cependant observé une faible association inverse entre le génotype homozygote Leu198Leu du SNP rs1050450 (*GPx1)* et le risque de cancer de la prostate. Jusqu'à récemment, ce polymorphisme n'était pas présent sur les puces GWAS les plus fréquemment utilisées, et il n'a été que récemment génotypé dans le cadre du projet HapMap et du projet 1000 génomes. Il est donc possible que cette association n'ait pas été détectée jusqu'à présent dans les précédentes études pangénomiques réalisées. Les approches gène-candidat précédemment mises en place ont fourni des résultats controversés sur cette association. Bien que certaines études aient

Figure 28 – Méta-analyse étudiant l'effet de rs1050450 sur le risque de cancer de la prostate

(a) Méta-analyse groupant notre étude avec les études déjà publiées en incluant l'étude de Kucukgergin[246]



(b) Méta-analyse groupant notre étude avec les études déjà publiées en excluant l'étude de Kucukgergin[246]



Le forest-plot des deux méta-analyses réalisées est représenté, accompagné des effectifs, des odds ratios de chaque étude ainsi que de l'odds ratio global, accompagnés de leurs intervalles de confiance à 95%. L'estimation de l'hétérogénéité entre les différentes études inclues est également présentée.

détecté une tendance à une association inverse entre le SNP rs1050450 et le risque de cancer de la prostate pour les porteurs homozygotes de l'allèle rare[247], d'autres n'ont observé aucun effet[244,245,248], ou encore une association augmentant le risque de cancer de la prostate[246]. Notre étude est la première à détecter une association significative au seuil de $\alpha = 5\%$, mais également la seule a avoir la puissance statistique suffisante pour détecter un tel effet. De plus, les résultats de la méta-analyse effectuée renforcent nos observations.

Aucune modification d'effet ou interaction significative entre la consommation d'alcool et le polymorphisme mitochondrial rs2853826 situé dans le gène *ND3* n'a été détectée dans le cadre de l'étude du risque de cancer du sein ou de la prostate. Cependant, les femmes porteuses de l'allèle G10398 et consommatrices d'alcool ont une durée de survie plus longue lorsque l'on étudie à la fois la survie totale et la survie spécifique au cancer du sein. Étrangement, les femmes porteuses de l'allèle A10398 et consommatrices d'alcool ont une durée de survie totale plus longue en comparaison de celles porteuses du même allèle mais ne consommant pas d'alcool. Une hypothèse explicative plausible serait l'existence d'une association inverse entre la consommation d'alcool et le risque de maladies cardiovasculaires[249,250]. Comme la plupart des femmes inclues

dans l'étude sont ménopausées (72%), et que le temps de suivi est long (en moyenne 8 années après le diagnostic), il est possible qu'un certain nombre de participantes à l'étude décèdent des suites d'une atteinte cardiovasculaire, et ce indépendemment du diagnostic de cancer du sein. Or, une consommation modérée d'alcool semblerait diminuer les risques de maladies cardiovasculaires. Le fait de ne plus observer de différences entre les courbes de survie spécifiques du cancer du sein tendrait à renforcer cette hypothèse. De plus, un certain nombre de traitements adjuvants tels que bevacizumab[251], taxane-anthracycline[252], and trastuzumab[253] sont associés avec des complications cardiaques lors du traitement du cancer du sein[254]. De plus, il a été montré qu'il existe des polymorphismes influant sur la cardiotoxicité de certains de ces traitements adjuvants[255,256]. De même, la radiothérapie peut augmenter le risque de maladie cardiovasculaire, et certains polymorphismes ont été identifiés comme pouvant augmenter la susceptibilité aux atteintes cardiovasculaires après radiothérapie[257]. Dans notre étude, le besoin de simplifier la définition des catégories d'exposition et d'homogénéiser les variables mesurant cette exposition entre les différentes cohortes du BPC3 inclues nous ont conduit à prendre en compte la consommation d'alcool de manière binaire, en catégorisant les consommateurs et les non consommateurs d'alcool. Une prise en compte plus détaillée de la consommation d'alcool, notamment en termes de quantité et du type d'alcool consommé pourrait permettre de mieux comprendre cette tendance. Il serait également très utile d'inclure dans ce type d'analyse les informations concernant les traitements administrés, information indisponible pour la grande majorité des femmes dans notre cas.

Ce travail de recherche a fait l'objet d'une publication en Janvier 2014, dans le journal *Free Radical Research*[258], dont l'intégralité est présentée en Chapitre III.2 .

Chapitre 2

Une approche phylogénétique originale de détection d'haplogroupes associés au risque de cancer du sein chez des porteuses de mutations sur BRCA1/2

L'axe de recherche développé dans cette seconde partie s'intéresse à l'étude d'une potentielle modification du risque de cancer du sein en fonction de l'haplogroupe mitochondrial chez des femmes porteuses d'une mutation pathogène sur *BRCA1* ou *BRCA2*.

Comme nous l'avons vu précédemment, la pénétrance du cancer du sein associée aux mutations pathogènes sur *BRCA1* et *BRCA2* est très élevée, de l'ordre de 40% à 85%. Rappelons également que parmi leurs diverses fonctions, les protéines encodées par ces deux gènes sont toutes deux impliquées dans la réparation des cassures simple-brin et double-brins, un type de dommage sur l'ADN pouvant être occasionné par l'exposition à divers agents génotoxiques, et en particulier par les espèces oxygénées réactives (ou ROS) générées au cours de la production d'ATP. La cascade de réactions biochimiques et métaboliques, aboutissant à la production d'ATP et à la génération de ROS, implique plusieurs protéines structurelles encodées par le génome mitochondrial. Les variations des gènes encodant ces protéines peuvent donc influencer l'efficacité de la production d'ATP, modifiant au passage l'équilibre de la synthèse des espèces oxygénées réactives. Dans ce contexte, on peut se demander dans quelle mesure des individus dont les mécanismes de réparation des dommages à l'ADN sont altérés voient leur risque de développer un cancer du sein modifié sous l'influence de variations localisées sur le génome mitochondrial.

Il existe de nombreux polymorphismes identifiés sur le génome mitochondrial, et d'innombrables combinaisons des allèles correspondants. Ces combinaisons d'allèles - ou haplotypes - peuvent être regroupées en haplogroupes selon certaines caractéristiques communes de leur séquence, chaque haplogroupe étant associé à un groupe ethnique donné. L'évolution la plus probable du génome mitochondrial ayant été reconstituée, on est capable de positionner les haplogroupes mitochondriaux les uns par rapport aux autres sur l'échelle de l'Évolution. La prise en compte d'informations phylogénétiques dans une étude d'association est une approche novatrice rendue possible par la relative simplicité du génome mitochondrial en comparaison du génome nucléaire.

L'étude présentée ici est basée sur l'analyse du génome mitochondrial d'un très grand nombre de femmes appartenant à l'étude CIMBA, dans le cadre du projet COGS. Les données analysées sont des données de génotypage d'un certain nombre de polymorphismes connus du génome mitochondrial. Un algorithme a été mis au point afin d'imputer l'haplogroupe mitochondrial à partir des données de génotypage en exploitant au maximum les informations disponibles.

Enfin, une approche phylogénétique originale couplée à des méthodes d'analyses statistiques classiques a été utilisée afin de détecter si, pour certains sous-groupes d'individus, le risque de cancer du sein est modifié sous l'influence de facteurs génétiques liés à la mitochondrie.

I. Matériels et Méthodes

I.1 L'étude COGS

L'étude *Collaborative Oncological Gene-environment Study*[259] (ou COGS), est un projet européen développé afin d'améliorer la compréhension de la susceptibilité génétique à trois types de cancer hormonodépendants : les cancers du sein, de l'ovaire et de la prostate. Ce projet implique plusieurs grands consortium internationaux : le *Breast Cancer Association Consortium*[260,261] (ou BCAC), le *Ovarian Cancer Association Consortium*[262,263](ou OCAC), le *Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome*[264,265] (ou PRACTICAL), et le *Consortium of Investigators of Modifiers of BRCA1/2*[266,267] (ou CIMBA). CIMBA est un groupe collaboratif de chercheurs travaillant sur les modificateurs génétiques du risque de cancer (principalement sein et ovaire) chez les femmes porteuses de mutations sur les gènes *BRCA1* et/ou *BRCA2*. En tant que groupe participant au projet COGS, plus de 200 000 SNPs dont 129 SNPs mitochondriaux ont été génotypés chez ces femmes pour lesquelles la recherche de mutations pathogènes sur les gènes *BRCA1/BRCA2* s'est révélée positive, et ce à l'aide de la puce iCOGS. La puce iCOGS est une puce de génotypage Illumina^{TM} Infinium à façon, conçue pour tester de manière ciblée et optimisée les variants génétiques liés aux cancers du sein, de l'ovaire et de la prostate.

I.2 Description des porteuses de mutation sur *BRCA1* et *BRCA2*

Les femmes inclues dans l'étude sont toutes âgées d'au moins 18 ans et portent une mutation pathogène sur les gènes *BRCA1* et/ou *BRCA2*. On connait leur année de naissance, leur âge au recrutement dans l'étude, leur âge au diagnostic du cancer du sein, leur ethnie, ainsi que la description des mutations détectées. Les analyses finales incluent 7432 femmes atteintes d'un cancer du sein et 7104 femmes non-affectées mutées sur *BRCA1*, ainsi que 3989 femmes atteintes d'un cancer du sein et 3689 femmes non-affectées mutées sur *BRCA2*. Toutes les analyses réalisées dans le cadre de cette étude ont été effectuées de manière indépendante sur la population de femmes porteuses de mutation sur *BRCA1* et sur la population de femmes porteuses de mutation sur *BRCA2* (respectivement abrégées *pop1* et *pop2* dans la suite du texte). Les femmes porteuses d'une mutation à la fois sur *BRCA1* et sur *RBCA2* sont inclues dans les deux analyses, sans traitement spécifique.

I.3 Génotypage et filtrage après contrôles qualité

129 SNPs ont été initialement génotypés à la fois pour les populations d'étude *pop1* et *pop2*. Le génotypage a été effectué sur la puce à façon Illumina^{TM} Infinium iCOGS. Les génotypes ont été établis en utilisant l'algorithme GenCall (propriété d'Illumina). Des analyses de contrôle

qualité ont été effectuées sur les données brutes. Les SNPs présentant les caractéristiques suivantes ont été exclus de la suite des analyses :

- les SNPs monoalléliques, dont la fréquence observée de l'allèle mineur dans nos données est nulle,
- les SNPs pour lesquels plus de 5% des génotypes sont manquants,
- les SNPs pour lesquels l'appel des génotypes a conduit à plus de 5% de génotypes hétérozygotes ; la mitochondrie étant haploïde, ces génotypes ne peuvent pas être considérés fiables.
- les SNPs étant annotés comme trialléliques ; la puce iCOGS permet d'étudier des marqueurs bialléliques uniquement. Pour les SNPs trialléliques, une partie de l'information est manquante, et les conclusions obtenues sur ces marqueurs peuvent donc être biaisées.
- les SNPs dont les sondes de génotypage s'hybrident sur le génome nucléaire. On exclut ces SNPs afin de s'assurer de la spécificité des éléments du génome mitochondrial génotypés.
- enfin, les SNPs représentant des mutations privées. Les mutations privées sont des mutations trop peu répandues dans la population générale pour être inclues dans l'arbre phylogénétique mitochondrial de référence (PhyloTree[147], voir Introduction, section III.3). Elles sont souvent spécifiques d'une ou de quelques familles uniquement.

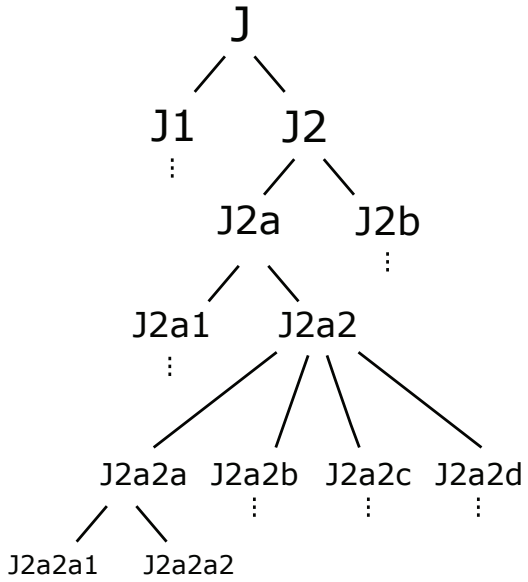
À l'issue de ces étapes de filtration, 93 SNPs pour la population *pop1*, et 92 SNPs pour la population *pop2* validant les critères de qualité requis ont été inclus dans la suite des analyses.

I.4 Arbre phylogénétique de référence du génome mitochondrial et haplogroupes mitochondriaux

Les analyses réalisées se basent sur l'arbre phylogénétique de référence du génome mitochondrial appelé PhyloTree[147] (Version 15). L'arbre complet est accessible à l'adresse <`http://www.phylotree.org/tree/main.htm`[148]>. L'arbre est enraciné par une séquence appelée *Reconstructed Sapiens Reference Sequence*, ou RSRS. Cette séquence a été établie par reconstruction de l'arbre évolutif mitochondrial[151] par parcimonie, une méthode couramment utilisée en phylogénie. Elle reconstruit l'arbre phylogénétique le plus probable en minimisant le nombre de changements nécessaires à introduire dans les séquences ancestrales afin d'établir une topologie cohérente avec les données observées. Ainsi la séquence RSRS est la séquence ancestrale commune à l'espèce humaine la plus probable à partir des données utilisées pour sa reconstruction. L'arbre obtenu inclut tous les haplogroupes connus, chacun d'entre eux étant individuellement défini par la liste de tous les SNPs qui se sont stabilisés dans la population pour une clade donnée de l'arbre. Chaque haplogroupe est ainsi entièrement caractérisé par une séquence de 16 569 bases. Cette séquence est le résultats de l'application de l'ensemble des substitutions correspondant aux SNPs qui définissent cet haplogroupe dans la séquence ancestrale RSRS.

Chaque haplogroupe correspond à un ensemble de clades, sous-clades, etc... La dénomination des haplogroupes mitochondriaux suit la logique suivante : chaque niveau de précision supplémentaire, correspondant à un niveau de profondeur de plus dans l'arbre évolutif, est représenté par un suffixe constitué à tour de rôle d'une lettre ou d'un nombre. Par exemple, comme illustré sur la figure 29, l'haplogroupe J est constitué des sous-clades J1 et J2 ; J2 de J2a et J2b ; J2a de J2a1 et J2a2 ; etc...

Figure 29 – Exemple de la clade issue de l'haplogroupe J



Prenons l'exemple de la séquence déposée sous GenBank ayant pour numéro d'accession FJ527772.2, qui correspond au génome mitochondrial d'un individu quelconque d'origine espagnole[268]. Cette séquence a pour haplogroupe principal H, mais elle est également classée H2, H2a, H2a5, H2a5a, H2a5a1 et H2a5a1a par niveau de précision successifs.

Les haplogroupes principaux, tels que les haplogroupes européens H, I, J, K, U, T, V, W et X, ont été définis bien avant que l'arbre évolutif ne soit reconstitué. C'est pourquoi, afin de conserver les notations préétablies, certains haplogroupes portent une dénomination qui échappe à la règle générale. Par exemple, la clade H et la clade V sont regroupées au sein de la clade HV7, étant elle-même une sous-clade de la clade HV.

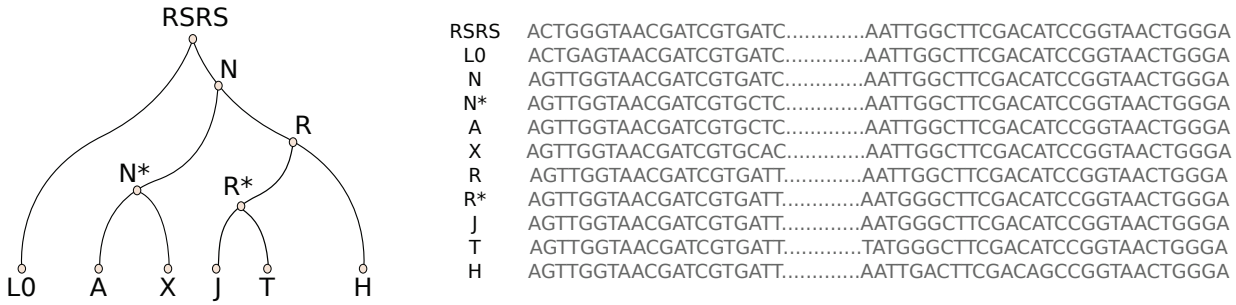
La classification d'une séquence dans l'arbre des haplogroupes revient à positionner cette séquence dans la branche correcte de l'arbre de référence. Plus les informations dont on dispose sont nombreuses, plus la précision dans la détermination de l'haplogroupe sera élevée. Concrètement, plus on a d'informations sur les génotypes de cette séquence, plus il sera aisé de positionner précisément la séquence dans l'arbre de référence, soit le plus près possible des feuilles de l'arbre.

I.5 Imputation des haplogroupes

La première étape de cette analyse consiste dans la mesure du possible à assigner à chaque individu son haplogroupe mitochondrial. L'haplogroupe le plus précis correspondant à une séquence de génome mitochondrial est parfaitement déterminé si on dispose de la séquence complète. Or, dans l'étude présentée ici, on ne dispose pas de la séquence dans son ensemble, mais uniquement d'un sous-ensemble de marqueurs ponctuels de cette séquence. Cet ensemble restreint de marqueurs permet donc de déterminer l'haplogroupe mitochondrial avec une précision qui ne sera pas maximale.

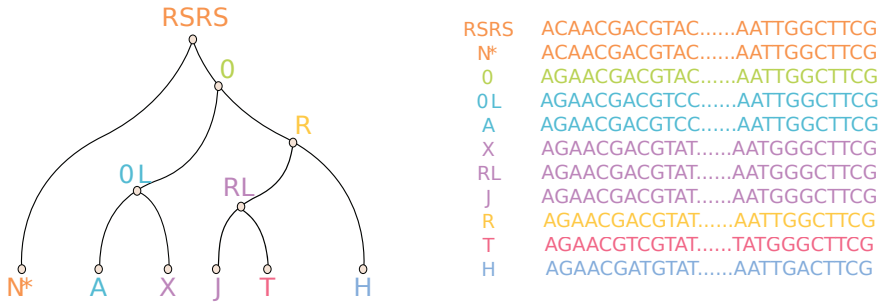
Tout d'abord, on travaille uniquement sur l'arbre de référence PhyloTree, en laissant nos propres données de côté. La séquence du génome mitochondrial peut être reconstruite à chaque noeud de l'arbre évolutif mitochondrial de référence, étant donné les substitutions qui se sont stabilisées dans la séquence RSRS au cours de l'Evolution. Chaque haplogroupe possède donc une séquence mitochondriale de pleine longueur qui lui est caractéristique.

Figure 30 – Étape 1 : Reconstitution des séquences pleine longueur aux noeud de l'arbre



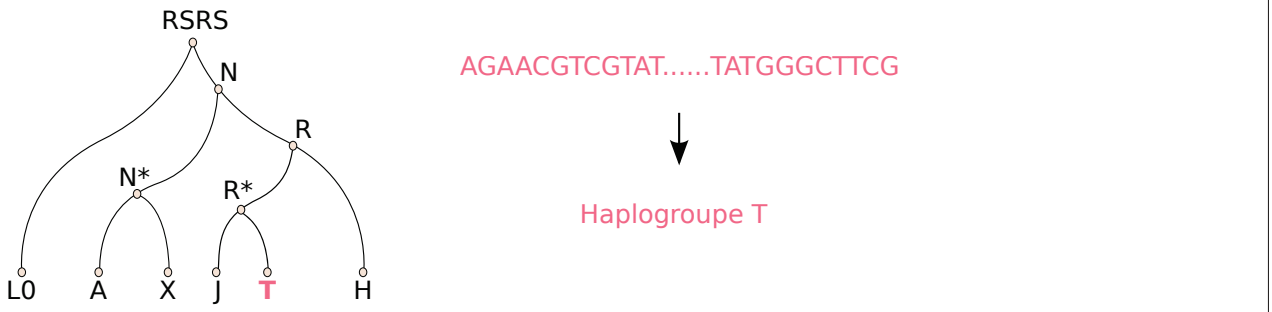
Cependant, nous ne disposons pas des séquences pleines longueur dans nos données, mais seulement de 93 et 92 SNPs pour les populations *pop1* et *pop2* respectivement. Pour les 7 864 noeuds de l'arbre évolutif mitochondrial de référence, on définit leur haplotype court, soit leur séquence haplotypique pleine longueur, mais restreinte aux loci disponibles.

Figure 31 – Étape 2 : Définition des haplotypes courts



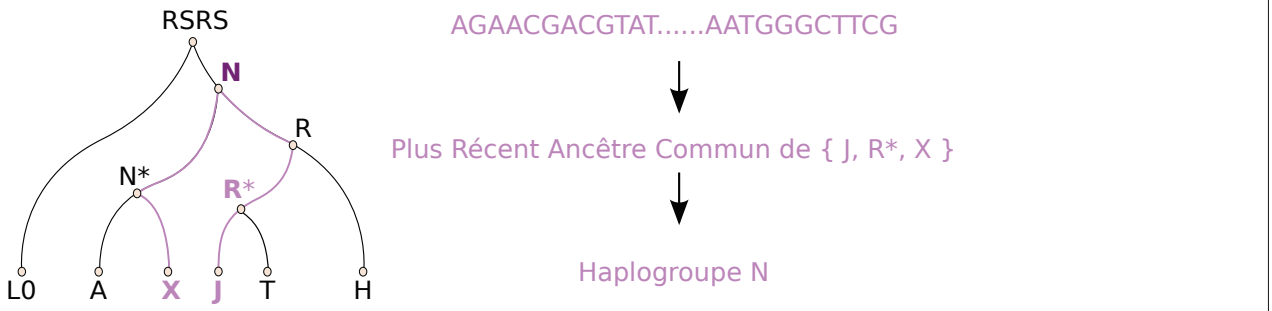
Certains haplotypes courts sont uniques, et une équivalence entre celui-ci et l'haplogroupe à ce noeud peut être faîte de manière simple et directe.

Figure 32 – Étape 3 : Assignation d'un haplogroupe à un haplotype court unique



Cependant la plupart du temps, étant donné le faible nombre de SNPs dont on dispose en comparaison de la longueur totale du génome mitochondrial, plusieurs haplogroupes portent le même haplotype court. Dans ce cas, on ne peut pas assigner un unique haplogroupe à cet haplotype court. Nous avons donc choisi d'assigner à cet haplotype court l'haplogroupe correspondant au plus récent ancêtre commun des haplogroupes partageant cet haplotype court. Ainsi, un haplogroupe est assigné à chaque haplotype court.

Figure 33 – Étape 4 : Assignation d'un haplogroupe à un haplotype multiple



Une fois que ces équivalences haplotype court $\Longleftrightarrow$ haplogroupe ont été établies, on reconstruit les haplotypes courts pour chaque individu de nos données, et on lui assigne l'haplogroupe correspondant.

La précision de la méthode d'inférence des haplogroupes décrite ci-dessus a été estimée en l'appliquant sur un ensemble de 630 séquences complètes de génomes mitochondriaux appartenant aux haplogroupes fréquents européens et caucasiens H, I, J, K, U, T, V, W et X. Ces

données proviennent de la base de données *Human Mitochondrial Genome Database*[269], et sont accessibles publiquement via l'url `http://www.genpat.uu.se/mtDB/`.

I.6 Détection d'association

Dans cette étude, nous utilisons une approche phylogénétique afin d'identifier les clades de l'arbre évolutif du génome mitochondrial différentiellement enrichies en individus affectés et non-affectés par rapport aux clades voisines. Nous utilisons un programme appelé ALTree[270,271] pour effectuer les tests d'association. ALTree - pour *Association detection and Localization of susceptibility sites using haplotype phylogenetic Trees* - est un algorithme effectuant des tests d'homogénéité emboîtés afin de comparer les distributions des individus affectés et non-affectés dans les clades d'un arbre phylogénétique donné. Le principe de cette méthode est présenté dans la figure 34. Il y a autant de tests effectués qu'il y a de niveaux dans l'arbre phylogénétique.

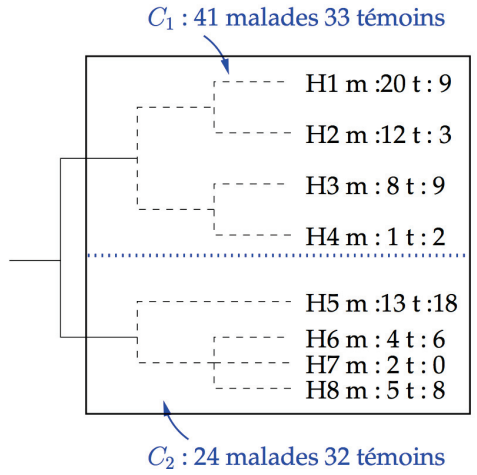
Une p-value est calculée à chaque test effectué selon différentes méthodes :
- par un test de X^2 classique dans le cas général,
- par un test exact de Fisher lorsque seulement deux clades sont comparées,
- par permutations lorsque plus de deux clades sont comparées et que certains effectifs sont faibles.

Dans ce dernier cas, avec un nombre de clades élevé et des effectifs faibles, la statistique du X^2_{obs} ne suit plus la distribution théorique du X^2. Des permutations sont requises afin d'évaluer dans quelle mesure la répartition observée des affectés/non-affectés entre les branches à ce niveau de l'arbre s'éloigne de la distribution théorique sous l'hypothèse nulle H_0, hypothèse selon laquelle il n'y a pas de différence de répartition entre les cas et les témoins au sein de l'arbre. Le principe des permutations repose sur le tirage d'un nombre élevé de rééchantilonnages, dans lesquels les statuts des individus sont aléatoirement permutés entre eux. Ainsi, pour chaque rééchantillonnage, le nombre global de cas et témoins ne change pas, mais leur répartition dans les clades de l'arbre est modifiée de manière aléatoire. Pour chacun de ces n rééchantilonnages, on calcule la statistique du X^2_i, i allant de 1 à n. La p-value correspond alors à la probabilité d'observer une valeur de la statistique qui soit supérieure à X^2_{obs} sous l'hypothèse H_0, c'est à dire suivant la distribution représentée par les X^2_i des n rééchantillonnages effectués. Concrètement, cette p-value s'obtient en déterminant la proportion p des n rééchantillonnages effectués tels que $X^2_i > X^2_{obs}$. Plus cette proportion p est faible, plus la probabilité d'observer X^2_{obs} sous l'hypothèse H_0 l'est également. Ainsi, pour $\alpha = 5\%$, on rejettera H_0 si $p < 0.05$.

Lorsque de grands arbres phylogénétiques sont étudiés, le nombre de tests peut rapidement augmenter. C'est pourquoi une procédure de correction des tests multiples est implémentée dans ALTree. Les tests emboîtés réalisés n'étant pas dépendants les uns des autres, une correction classique du type Bonferroni serait beaucoup trop conservatrice, et on perdrait beaucoup trop de puissance statistique. Une procédure de correction des tests multiples adaptées aux tests emboîtés[273] fonctionnant par permutations a donc été préférée pour ALTree. L'objectif de cette méthode étant de détecter une association globale au niveau de l'arbre, seule la plus faible p-value obtenue pour tous les tests effectués - soit la plus significative dans un arbre donné - est corrigée. Dans cette étude 1000 permutations sont effectuées pour chaque arbre.

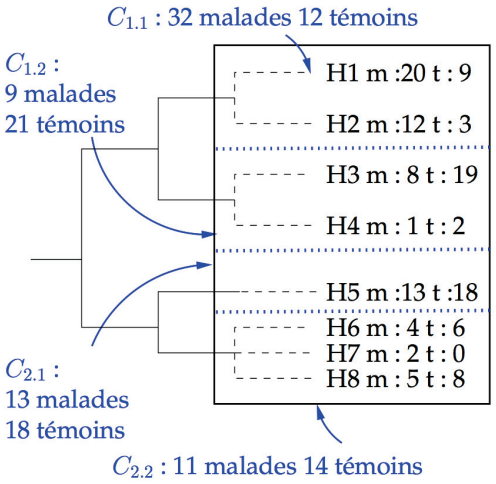
FIGURE 34 – Description de l'analyse emboîtée lors du test d'association. D'après C. Bardel[272]

(a) Niveau 1



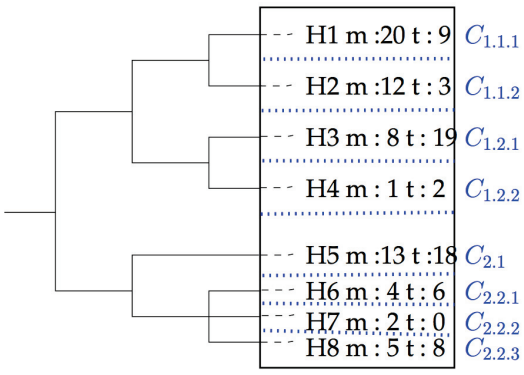
Au niveau 1 (première division dans l'arbre), un test d'homogénéité de la distribution des affectés et des non-affectés entre les clades C_1 et C_2 est réalisé (1 degré de liberté). La p-value de ce test est de 0,16.

(b) Niveau 2



Au niveau 2, un test d'homogénéité de la distribution des affectés et des non-affectés entre les clades $C_{1.1}$, $C_{1.2}$, $C_{2.1}$ et $C_{2.2}$ — qui sont les descendants des clades C_1 and C_2 — est réalisé (3 degrés de liberté). La p-value de ce test est de 0,0019.

(c) Niveau 3



Au niveau 3, un test d'homogénéité de la distribution des affectés et des non-affectés entre les clades $C_{1.1.1}$, $C_{1.1.2}$, $C_{1.2.1}$, $C_{1.2.2}$ $C_{2.1}$, $C_{2.2.1}$, $C_{2.2.2}$ et $C_{2.2.3}$ est réalisé (7 degrés de liberté). La p-value de ce test est de 0,057.

I.7 Gestion de la dépendance génétique

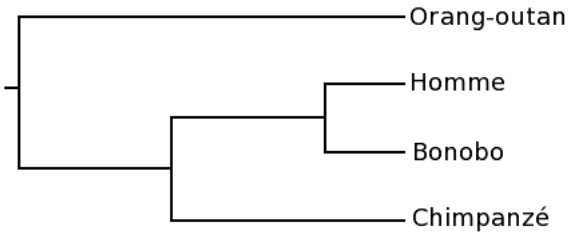
ALTree effectue des tests d'homogénéité pour détecter des différences d'enrichissement en individus affectés et non-affectés entre les clades d'un arbre phylogénétique. Ce type de test ne peut être légitimement effectué que si les données sont indépendantes. Or, dans notre jeu de données, certains individus appartiennent à la même famille. C'est pourquoi nous construisons un jeu de données constitué d'individus indépendants du point de vue du génome mitochondrial en sélectionnant aléatoirement un individu parmi tous ceux appartenant à la même famille ET présentant un haplotype court identique. En effet, bien que n'étant pas indépendants génétiquement si on considère leur génome nucléaire, deux individus apparentés mais ayant un génome mitochondrial différent (certaines cousines par exemple) sont génétiquement indépendants si on ne considère que leur génome mitochondrial. L'étude présentée ici ne s'intéresse qu'aux variants du génome de la mitochondrie, et non aux éventuelles interactions entre le génome nucléaire et le génome mitochondrial. Il est donc légitime de ne considérer l'indépendance des individus inclus que du point de vue du génome mitochondrial. Pour tenir compte de l'entière variabilité de nos données, le rééchantillonnage présenté est effectué 1000 fois, générant ainsi 1000 jeux de données constitués d'individus indépendants. La topologie des phylogénies reconstruites par la suite (c'est à dire la forme des arbres reconstruits) sera identique entre tous les rééchantilonnages, puisqu'on sélectionne aléatoirement des individus parmi ceux ayant un haplotype court identique. Seuls les effectifs des individus affectés ou non-affectés sont susceptibles de changer dans un arbre donné. Les analyses d'association sont appliquées sur chaque jeu de données de manière indépendante. Les valeurs des résultats obtenus pour tous les rééchantilonnages sont ensuite moyennées pour obtenir des résultats globaux.

I.8 Reconstruction des caractères aux noeuds ancestraux

ALTree comprend également un algorithme de localisation des sites de susceptibilité dans le cas où une association est détectée. Cependant, avant de pouvoir lancer l'étape de localisation, les séquences ancestrales doivent être reconstruites. Autrement dit, les haplotypes courts doivent être reconstruits pour tous les noeuds de l'arbre qui ne sont pas des noeuds terminaux. Nous avons utilisé le programme PAML[274] pour effectuer cette reconstruction par maximum de vraisemblance. Le modèle phylogénétique utilisé est le modèle *General Time-Reversible* (communément désigné par GTR ou REV). La reconstruction nécessite d'avoir à disposition un groupe externe, appelé aussi un *outgroup*. Un *outgroup* est par définition la séquence d'une espèce proche du reste des espèces constituant une phylogénie donnée, mais plus éloignée de chacune d'elles que n'importe quelle paire d'espèces au sein de cette phylogénie (Figure 35).

PAML ayant besoin d'un *outgroup* pour effectuer la reconstruction des caractères ancestraux, la séquence du génome mitochondrial de *homo neanderthalensis* (Homme de néandertal) ayant pour identifiant GenBank `NC_011137.1` a été choisie. Après alignement local contre la séquence RSRS, les allèles correspondant aux positions homologues des SNPs dont on dispose sont extraits du génome mitochondrial néandertalien afin de reconstruire l'haplotype court de cette espèce. Cet haplotype court est utilisé comme *outgroup*.

Figure 35 – Exemple d'outgroup



L'orang-outan est plus éloigné de l'homme, du bonobo et du chimpanzé que chacune de ces trois espèces les unes par rapport aux autres. Ainsi, l'orang-outan constitue un outgroup approprié pour cette phylogénie.

I.9 Localisation des sites de susceptibilité

ALTree permet également d'identifier quels sites, c'est à dire quels SNPs constituant les haplotypes testés sont les plus susceptibles d'être impliqués dans l'association détectée, si association il y a. Pour cela, pour chaque haplotype court observé, une extension d'ALTree appelée `altree-add-S` ajoute à la séquence de l'haplotype court un caractère supplémentaire appelé `S`. Ce caractère `S` représente le statut (affecté ou non-affecté) associé à cet haplotype court ; les individus portant cet haplotype sont-ils plus fréquemment affectés ou non-affectés ?

Le caractère `S` est déterminé en fonction de la proportion de cas et de témoins chez les individus présentant un haplotype donné, selon les règles suivantes, p_h étant la proportion d'individus malades présentant l'haplotype h, p_0 la proportion globale de malades chez tous les individus, n_h le nombre d'individus présentant l'haplotype h, et ϵ un paramètre fixé par l'utilisateur dont la valeur par défaut est 1, influant sur la souplesse de la détermination du caractère `S` :

- Si $p_h < p_0 - \epsilon.\sqrt{\frac{p_h.(1-p_h)}{n_h}}$, `S` est codé **C** : la majorité des individus présentant l'haplotype h sont sains.

- Si $p_h > p_0 + \epsilon.\sqrt{\frac{p_h.(1-p_h)}{n_h}}$, `S` est codé **G** : la majorité des individus présentant l'haplotype h sont malades.

- Sinon, `S` est codé **?** : les individus présentant l'haplotype h ne sont majoritairement ni sains ni malades.

Les états de caractère aux noeuds ancestraux sont reconstruits en incluant `S` de manière à pouvoir déterminer au niveau de quelles branches de l'arbre évolutif il est le plus probable que des changements du caractère `S` aient eu lieu. ALTree effectue alors la localisation des sites de susceptibilité en calculant un indice de coévolution. Cet indice a pour but de détecter les sites de la séquence haplotypique qui varient en même temps que le caractère S, et ce dans les deux sens possibles de substitution. Les sites pour lesquels les changements observés sont le plus corrélés avec ceux du caractère `S` sont les sites de susceptibilité les plus probables.

I.10 Sélection des sous-clades

Les analyses ont été effectuées sur l'arbre évolutif complet. Cependant, plus il y a d'haplogroupes à un niveau donné de l'arbre, plus le nombre de degrés de liberté des tests réalisés est élevé, moins la puissance statistique des tests sera élevée. C'est pourquoi, les analyses ont également été effectuées sur des sous-clades de l'arbre. Les sous-clades ont été choisies en fonction du nombre d'individus et d'haplogroupes qu'elles regroupent, afin de maximiser la puissance statistique. Les sous-clades choisies ainsi que les effectifs associés sont présentés dans la Table 6.

Sous-clade	**Porteurs de mutation sur *BRCA1***	**Porteurs de mutation sur *BRCA2***
U8	1458	863
T	1243	651
J	1270	630
J1	1043	513
H	3706	1967
H1	582	337
U5	868	458
X1'2'3	221	103
K1a	608	364

Table 6 – Effectifs des participantes dans chaque sous-clade sélectionnée

I.11 Analyses statistiques

Dans le cas où une différence d'enrichissement en affectés/non affectés est détectée par ALTree, nous avons quantifié l'effet associé à l'aide d'une régression de Cox pondérée. Pour chaque modèle construit, les données ont été restreintes à la sous-clade d'intérêt. D'autre part, le taux d'incidence du cancer du sein dépend du gène porteur de la mutation pathogène (*BRCA1* ou *BRCA2*), et évolue en fonction de l'âge de l'individu. C'est pourquoi une méthode de pondération prenant en compte ces deux éléments[275] a été appliquée dans le modèle. Enfin, la dépendance familiale entre individus apparentés a été prise en compte en utilisant une matrice corrigée dite robuste à l'aide d'un estimateur de type « sandwich » (package R *survival*, fonction *cluster()*).

I.12 Ressources computationnelles et temporelles

Deux étapes de cette suite d'analyses sont particulièrement longues : la première est la procédure des 1000 permutations mise en place afin d'estimer les p-values corrigées des tests d'homogénéité, la seconde est la nécessité d'appliquer l'ensemble des analyses sur les 1000 rééchantillonnages à cause de l'apparentement entre certains individus. Bien qu'ALTree soit codé en Perl, la procédure de permutations est elle implémentée en C et parallélisée sur les coeurs d'une même machine. Approximativement, ces analyses auront requis environ 36 heures de calcul parallèle sur une machine de 24 coeurs avec 256 Gigaoctets de RAM.

II. Résultats

II.1 Inférence des haplogroupes individuels

La première étape de cette analyse est l'inférence des haplogroupes pour chaque individu inclus dans l'étude. La Table 7 récapitule les effectifs et les fréquences des haplogroupes inférés pour les populations *pop1* et *pop2*. Pour les porteurs de mutations sur *BRCA1*, 489 haplotypes courts distincts ont été reconstruits à partir des 93 loci disponibles dans nos données. Seuls 162 de ces 489 haplotypes courts sont observés parmi les haplotypes courts théoriques reconstruits dans l'arbre de référence. Cependant, ces 162 haplotypes représentent 13315 sur les 14536 individus de *pop1* ; il a donc été possible d'assigner un haplogroupe à 91.6% des individus mutés sur *BRCA1*. De même pour les individus mutés sur *BRCA2*, 350 haplotypes courts d'une longueur de 92 SNPs ont été reconstruits. Seulement 139 sur 350 sont observés parmi les haplotypes courts théoriques reconstruits dans l'arbre évolutif mitochondrial. Ces 139 haplotypes représentent 6996 des 7678 individus mutés sur *BRCA2*, il a donc été possible d'assigner un haplogroupe à 91.1% des effectifs pour cette population.

Table 7 – Effectifs et Fréquence (%) des haplogroupes imputés par population d'étude

Haplogroupe	**Effectifs (*pop1*)**	**%**	**Effectifs (*pop2*)**	**%**
A	5	0.038	2	0.029
A2	4	0.03	2	0.029
A2e	2	0.015	1	0.014
A2k1	1	0.008	-	-
A4b	2	0.015	5	0.071
B4a1	2	0.015	3	0.043
B4a1a	-	-	1	0.014
B4b1a3	1	0.008	-	-
B4b'd'e'j	2	0.015	4	0.057
B4c	1	0.008	-	-
C	20	0.15	3	0.043
C1b3	1	0.008	-	-
C4	3	0.023	4	0.057
D4	8	0.06	5	0.071
F1b1	3	0.023	-	-
F3b1	-	-	1	0.014
H	1799	13.511	912 1	3.036
H13a2b2	6	0.045	9	0.129
H13b1	9	0.068	8	0.114
H1ag1b	1	0.008	1	0.014
H1at	12	0.09	15	0.214
H1au1b	1	0.008	-	-
H1b1e	5	0.038	8	0.114
H1bh	3	0.023	3	0.043
H1bm	5	0.038	1	0.014
H1c	348	2.614	187	2.673
H1c14	-	-	2	0.029
H1e	164	1.232	94	1.344
H1j2	14	0.105	7	0.1
H1n1b	24	0.18	18	0.257
H1v1b	5	0.038	1	0.014
H20b	1	0.008	-	-
H2a	262	1.968	139	1.987
H2a1a	22	0.165	7	0.1
H2a2b4	2	0.015	3	0.043
H2a5	15	0.113	15	0.214
H3	544	4.086	307	4.388
H3b1	21	0.158	17	0.243

Haplogroupe	**Effectifs (*pop1*)**	**%**	**Effectifs (*pop2*)**	**%**
H3q	6	0.045	1	0.014
H3s	7	0.053	8	0.114
H41a	41	0.308	30	0.429
H4a	22	0.165	16	0.229
H5a	317	2.381	135	1.93
H5a1c1	5	0.038	1	0.014
H5q	1	0.008	1	0.014
H65a	15	0.113	6	0.086
H6a1a1	12	0.09	7	0.1
H6a1b3a	8	0.06	5	0.071
H7d3	9	0.068	3	0.043
HV	2870	21.555	1447	20.683
HV2a	8	0.06	11	0.157
HV7	2	0.015	2	0.029
J1	6	0.045	5	0.071
J1b	10	0.075	3	0.043
J1b1	8	0.06	1	0.014
J1b1a	112	0.841	45	0.643
J1c	728	5.468	375	5.36
J1c1	100	0.751	51	0.729
J1c1b1a	17	0.128	6	0.086
J1c2b1	2	0.015	2	0.029
J1c2f	13	0.098	2	0.029
J1c5f	7	0.053	8	0.114
J1c8	40	0.3	15	0.214
J2a	151	1.134	72	1.029
J2b1	76	0.571	45	0.643
K	1	0.008	-	-
K1	456	3.425	285	4.074
K1a12	3	0.023	6	0.086
K1a17a	4	0.03	-	-
K1a1b	393	2.952	221	3.159
K1a1b2a1	16	0.12	10	0.143
K1a2c	2	0.015	2	0.029
K1a4	67	0.503	25	0.357
K1a4a1	122	0.916	99	1.415
K1a4a1d	1	0.008	1	0.014
K1c1 7	7	0.578	31	0.443
K1c1d	5	0.038	2	0.029
K2	264	1.983	152	2.173
K2a3a1	5	0.038	1	0.014
L0	-	-	5	0.071
L0a	1	0.008	1	0.014
L0a2	3	0.023	-	-
L0d3	2	0.015	-	-
L1b1	17	0.128	1	0.014
L1c1d	-	-	2	0.029
L2	1	0.008	-	-
L2a1	31	0.233	15	0.214
L2a1a	1	0.008	-	-
L2a1c	12	0.09	1	0.014
L2a1n	-	-	2	0.029
L3	2	0.015	5	0.071
L3'4	2	0.015	7	0.1
L3d	1	0.008	1	0.014
L3e1c	2	0.015	2	0.029
L3e'i'k'x	7	0.053	3	0.043
M	40	0.3 3	7	0.529
M11	1	0.008	-	-
M33a1b	1	0.008	-	-
M8a	1	0.008	-	-
N	13	0.098	3	0.043
N1	18	0.135	21	0.3
N1a	2	0.015	2	0.029
N1a1	37	0.278	14	0.2
N1a1b	89	0.668	18	0.257
N1b1b	114	0.856	62	0.886

Haplogroupe	**Effectifs (*pop1*)**	**%**	**Effectifs (*pop2*)**	**%**
N2	3	0.023	-	-
N9a1	1	0.008	-	-
R	57	0.428	35	0.5
R0a2b	3	0.023	1	0.014
R1	10	0.075	10	0.143
R1a1a	1	0.008	-	-
R2	-	-	2	0.029
T	70	0.526	38	0.543
T1a1	194	1.457	119	1.701
T1a1e	1	0.008	-	-
T2	359	2.696	180	2.573
T2b	562	4.221	277	3.959
T2b6a	3	0.023	3	0.043
T2b7a	24	0.18	13	0.186
T2c1c1	6	0.045	4	0.057
T2c1d2	1	0.008	-	-
T2d1	4	0.03	5	0.071
T2g	19	0.143	12	0.172
U	586	4.401	283	4.045
U1c	1	0.008	9	0.129
U2d2	7	0.053	3	0.043
U2e	168	1.262	66	0.943
U2e1b1	13	0.098	4	0.057
U3a1	85	0.638	48	0.686
U3b1b	9	0.068	-	-
U4b1a2	12	0.09	6	0.086
U4d2	8	0.06	-	-
U5a1	254	1.908	124	1.772
U5a1a1	150	1.127	68	0.972
U5a1a2b	6	0.045	5	0.071
U5a1a2b1	3	0.023	-	-
U5a1b1c	10	0.075	1	0.014
U5a2a1a	2	0.015	-	-
U5b	425	3.192	252	3.602
U5b2a1a2	11	0.083	6	0.086
U5b3g	7	0.053	2	0.029
U6	26	0.195	20	0.286
U7	-	-	17	0.243
U8a	36	0.27	23	0.329
U8b	6	0.045	5	0.071
V3b	3	0.023	1	0.014
W	179	1.344	98	1.401
W3a1	38	0.285	22	0.314
X	6	0.045	6	0.086
X2	89	0.668	46	0.658
X2a2	1	0.008	-	-
X2b7	15	0.113	9	0.129
X2b'd	110	0.826	47	0.672
X2g	2	0.015	-	-
X3	4	0.03	1	0.014

II.2 Précision de la méthode d'imputation des haplogroupes

Notre méthode d'inférence des haplogroupes a été testée sur un set de 630 séquences de génomes mitochondriaux dont l'haplogroupe principal est connu. Toutes ces séquences appartiennent à un des haplogroupes européens et caucasiens fréquents parmi H, I, J, K, U, T, V, W et X. De la même manière que pour environ 9% des individus issus de notre jeu de données expérimentales, la méthode employée n'a pas été capable d'inférer un haplogroupe pour 42 séquences, soit 6.7% des séquences testées. Ainsi, pour ces 42 séquences, les combinaisons observées des 92 SNPs d'intérêt ne sont pas retrouvées dans l'arbre de référence PhyloTree. L'origine ethnique des individus correspondants aux séquences pour lesquelles l'haplogroupe n'a pas pu être imputé est présentée en Table 8.

Table 8 – Origine ethnique des individus dont le génome mitochondrial n'a pas pu être rattaché à un haplogroupe dans le jeu de séquences de test

Origine ethnique	**Nombre d'individus**
Américains, Européens, Caucasiens	16
Sardes et Siciliens (Italie)	8
Finnois (Finlande)	7
Andalous et Castillans (Espagne)	5
Arméniens	1
Géorgiens	1
Cherkessk (Russie occidentale)	1
Néo-zélandais	1
Autres	2

Pour les 93.3% des séquences pour lesquelles l'imputation d'un haplogroupe a abouti, nous avons analysé si la méthode utilisée avait imputé avec succès l'haplogroupe principal (Soit un des haplogroupes parmi H, I, J, K, U, T, V, W et X). La précision de la méthode employée a été estimée à partir de ces résultats, soit la proportion de succès de l'imputation de l'haplogroupe principal pour les 630 séquences testées. Les résultats de validation de l'imputation des haplogroupes sont présentés dans la Table 9. Pour ce jeu de données test, notre méthode a été capable de retrouver l'haplogroupe référencé pour 82% des séquences testées. La précision atteint 100% pour les séquences appartenant aux haplogroupes I, J, K, T, U, W et X. Elle est de 70% pour les séquences appartenant à l'haplogroupe H, et tombe à 3% pour les séquences appartenant à l'haplogroupe V. Pour les 30% des séquences H et les 97 % des séquences V pour lesquelles la méthode n'a pas retrouvé l'haplogroupe correct, l'haplogroupe imputé était HV ou HV7.

Table 9 – Résultats de l'imputation et précision de la méthode en fonction de l'haplogroupe principal et de certaines sous-clades

Haplogroupe	Nb initial	Nb imputées	Taux d'imputation	Succès	Précision
H	155	143	92%	100	70%
I	18	17	94%	17	100%
J	74	69	93%	69	100%
K	50	46	92%	46	100%
T	72	69	96%	69	100%
U	143	131	91%	131	100%
V	64	62	97%	2	3%
W	42	39	93%	39	100%
X	12	12	100%	12	100%
Total	*630*	*588*	*93%*	*485*	*82%*
T1	24	23	96%	18	78%
T1a1	19	18	95%	18	100%
T2	48	46	96%	46	100%

Nb initial : Nombre de séquences initial pour cet haplogroupe
Nb imputées : Nombre de séquences pour lesquelles un haplogroupe a été imputé
Taux d'imputation : proportion de séquences de cet haplogroupe pour lesquelles l'imputation a abouti
Succès : Nombre de séquences pour lesquelles l'haplogroupe imputé est correct
Précision : Nombre de succès rapporté au nombre total de séquences imputées pour cet haplogroupe

II.3 Recherche d'associations

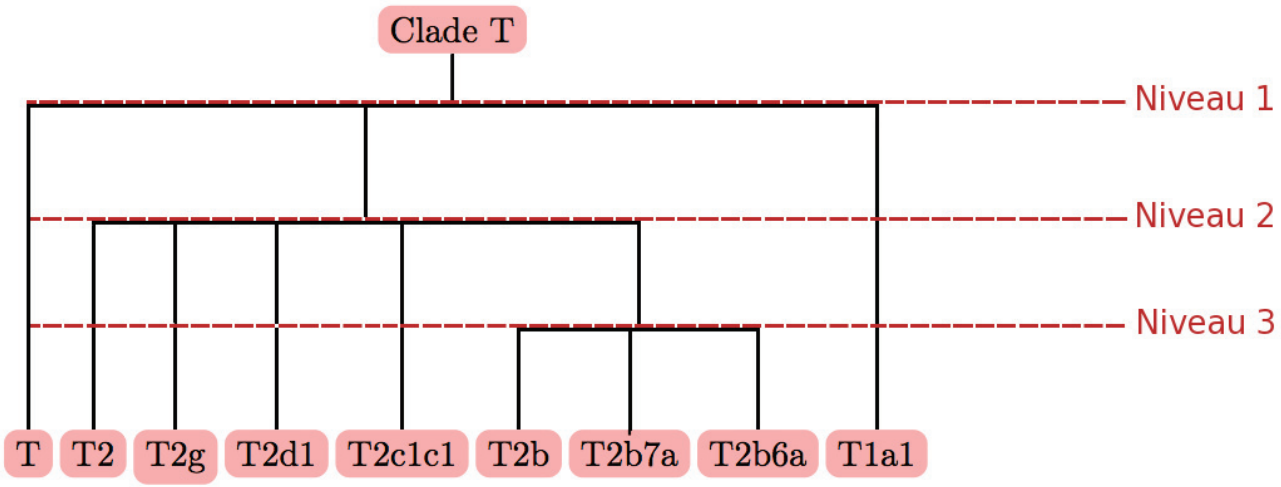
Une fois les haplogroupes imputés pour les individus de notre étude, la détection d'association est effectuée avec ALTree. Pour chacune des deux populations étudiées, et pour chaque analyse effectuée sur l'arbre évolutif complet comme sur chaque sous-clade, les p-values corrigées sont extraites pour chacun des 1000 rééchantilonnages effectués afin de tenir compte de l'apparentement entre certains individus. Elles sont ensuite moyennées (voir Table 10). La seule p-value corrigée moyenne significative au seuil $\alpha = 5\%$ est celle obtenue pour la population des individus porteurs d'une mutation sur *BRCA2* pour la sous-clade T (notée T*).

Table 10 – P-values corrigées des tests d'association obtenues avec ALTree par population d'étude et par sous-clade, moyennées sur les 1000 rééchantilonnages effectués

Sous-clade	*pop1*	*pop2*
Full	0.8298671	0.680985
U8	0.1457532	0.6260519
T	0.2854815	**0.0402038**
J	0.7175275	0.1115694
J1	0.6214585	0.1491129
H	0.7474476	0.9302557
H1	0.2677572	0.8035485
U5	0.8293806	0.7474615
X1'2'3	0.4155115	0.6288012
K1a	0.1701149	0.1617493

Quelque soit la sous-clade testée, l'arbre utilisé comme support à la détection d'association contient plusieurs niveaux, et plusieurs tests d'homogénéité sont donc réalisés. Seule la p-value la plus significative est corrigée par la procédure employée. On ne sait donc pas d'emblée à quel niveau de l'arbre de la sous-clade T la p-value significative correspond. L'arbre phylogénétique de la sous-clade T contient 3 niveaux (Figure 36).

Figure 36 – Arbre phylogénétique de la sous-clade T utilisé dans la recherche d'association avec ALTree



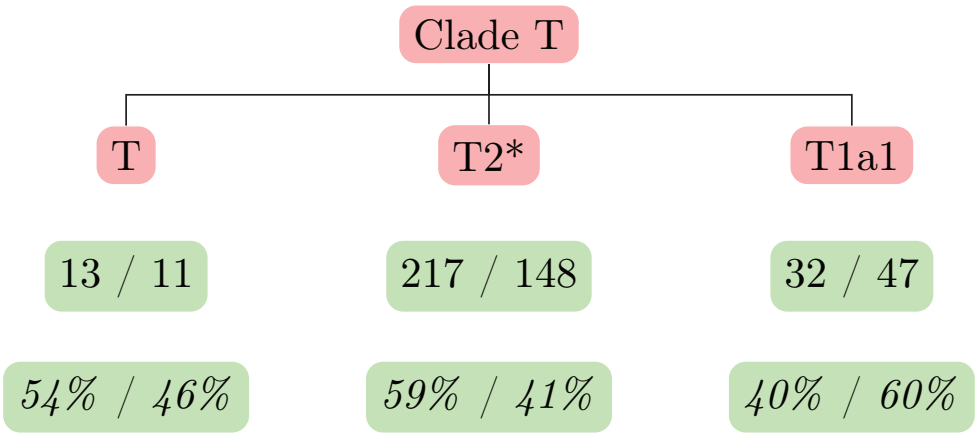
L'observation des p-values non-corrigées de la sous-clade T (Table 11) nous indique que la p-value corrigée significative correspond au premier niveau de l'arbre évolutif de la sous-clade T, c'est donc au premier niveau de l'arbre qu'une différence d'enrichissement en individus affectés et non-affectés a été détectée.

Table 11 – Moyennes des p-values non-corrigées par niveau de l'arbre phylogénétique de la sous-clade T chez les porteurs de mutation sur *BRCA2*

Niveau	Degrés de Liberté	p-value
1	2	0.02141039
2	6	0.14355900
3	8	0.22249700

La figure 37 représente les effectifs moyens des individus affectés et non-affectés au premier niveau de l'arbre. Ainsi, l'ensemble de ces observations suggère que la sous-clade T1a1 est de manière statistiquement significative moins enrichie en individus affectés que les sous-clades voisines T et T2.

Figure 37 – Représentation et effectifs du 1er niveau de l'arbre de la sous-clade T



T2* représente l'ensemble de la sous-clade T2. Les effectifs représentés sur fond vert correspondent respectivement au nombre d'individus affectés et non-affectés, moyennés sur l'ensemble des rééchantillonnages effectués. En dessous des effectifs sont représentés les pourcentages correspondants aux effectifs.

II.4 Résultats de localisation

La localisation des sites de susceptibilité les plus probables a été effectuée avec ALTree pour la sous-clade T chez les femmes mutées sur *BRCA2*. Les indices de coévolution calculés au cours de cette procédure sont présentés dans la Table 13. Plus l'indice de coévolution est élevé, plus la probabilité que ce site soit impliqué dans l'association détectée l'est également. Parmi les 92 sites constituant les haplotypes, les sites numérotés 44, 57 et 72 se distinguent. La table 12 contient des informations supplémentaires sur ces trois sites.

Table 12 – Description des SNPs identifiés comme potentiels sites de susceptibilité par ALTree.

Site	SNP	Position	Sens	Indice de Coévolution	Allèle Majeur	Allèle Mineur	Fréquence de l'allèle mineur*
44	MitoT9900C	9899	T→C	0.390	T	C	0.0163
57	rs41544217	11812	G→A	0.324	A	G	0.0709
72	rs28359178	13708	G→A	0.318	G	A	0.1106

*Fréquence calculée sur *pop2*

II.5 Quantification de l'effet détecté

Nous avons estimé le risque de cancer du sein des individus T1a1 en comparaison à celui des individus porteurs d'un autre haplogroupe de la sous-clade T dans la population des porteurs de mutation sur *BRCA2*. Nous avons évalué un *Hazard-Ratio* de 0.54, avec pour intervalle de confiance [0.34 - 0.87](p-value = 0.00914). Nous avons également comparé l'haplogroupe T1a1 avec les autres haplogroupes de la sous-clade T ainsi que l'haplogroupe H, le plus fréquent dans la population générale européenne. Le *Hazard-Ratio* correspondant est de 0.62, avec pour intervalle de confiance [0.40 - 0.98] (p-value = 0.0413).

III. Discussion

Dans cette étude, nous avons utilisé une méthode analytique originale basée sur la phylogénie, couplée des analyses d'épidémiologie moléculaire plus classiques, afin de détecter si certains haplogroupes mitochondriaux sont différentiellement enrichis en individus affectés et non-affectés parmi les femmes porteuses de mutations sur les gènes *BRCA1* et *BRCA2*. Nous avons réussi à imputer les haplogroupes des individus inclus dans l'étude avec succès pour plus de 90% d'entre eux. À l'issue de l'imputation des haplogroupes, l'utilisation d'ALTree a permis d'identifier la branche T1a1 appartenant à l'haplogroupe T comme différentiellement enrichie en individus affectés ou non-affectés parmi les porteurs de mutation sur *BRCA2*, alors qu'aucune différence d'enrichissement n'a été observée parmi les porteurs de mutation sur *BRCA1*.

Table 13 – Indices de coévolution calculés pour tous les sites non-monomorphiques constituant les haplotypes courts de la sous-clade T

SNP	nº Site	Sens	Indice de coévolution
MitoG3012A	site 4	A→G	-0.270
		G→A	-0.101
MitoC3595T	site 9	C→T	-0.101
		T→C	-0.270
MitoG3916A	site 11	A→G	-0.270
		G→A	-0.101
rs1599988	site 14	C→T	-0.101
		T→C	-0.270
MitoA4918G	site 19	A→G	-0.270
		G→A	-0.101
rs3021087	site 21	A→G	-0.232
		G→A	-0.101
MitoG5461A	site 23	A→G	-0.270
		G→A	-0.101
rs2298011	site 41	A→G	-0.101
		G→A	-0.270
MitoT9900C	site 44	C→T	-0.270
		T→C	**0.390**
MitoG10399A	site 51	A→G	-0.101
		G→A	-0.270
MitoT10464C	site 52	C→T	-0.101
		T→C	-0.270
rs41544217	site 57	A→G	-0.101
		G→A	**0.324**
rs2853497	site 60	A→G	-0.270
		G→A	-0.101
MitoT12706C	site 64	C→T	-0.101
		T→C	-0.270
rs28477492	site 66	C→T	-0.270
		T→C	-0.101
MitoA13106G	site 67	A→G	-0.101
		G→A	-0.270
rs28604589	site 69	A→G	-0.101
		G→A	-0.270
rs28359178	site 72	A→G	-0.270
		G→A	**0.318**
rs28357676	site 80	A→G	-0.270
		G→A	-0.101
MitoT14799C	site 84	C→T	-0.270
		T→C	-0.101
rs28357372	site 88	A→G	-0.270
		G→A	-0.101

Plus l'indice de coévolution d'un site est élevé, plus la probabilité que ce site soit impliqué dans l'association détectée l'est également. Les sites en position 44, 57 et 72 dans la séquence de l'haplotype court se démarquent : ce sont les seuls positifs.

L'haplogroupe T représente une proportion variable de la population générale selon l'ethnie, allant de 4% chez les Africains jusqu'à 11% en Europe de l'est et chez les Caucasiens. Dans notre jeu de données l'ensemble de la clade T représente 9.34% des porteurs de mutation sur *BRCA1*, et 9.30% des porteurs de mutation sur *BRCA2*. Grâce à sa fonction de localisation, ALTree a de plus identifié 3 sites de susceptibilité potentiels au sein du génome mitochondrial.

Dans cette étude, nous avons analysé dans quelle mesure la variabilité du génome mitochondrial modifie le risque de cancer du sein conféré par la mutation pathogène portée sur les gènes *BRCA1/2*. À l'heure actuelle, une large proportion de l'héritabilité du cancer du sein reste inexpliquée[276]. Différentes méthodes existent afin d'étudier la susceptibilité génétique à une maladie, comme les études de liaison génétiques (qui ont conduit à l'identification des gènes *BRCA1* et *BRCA2*), ou les études d'association pangénomiques, plus connues sous l'acronyme GWAS - pour *Genome Wide Association Studies*. Cependant, les études de liaison génétique ne s'appliquent pas au génome mitochondrial. En outre, les puces commerciales utilisées en GWAS pour le génotypage ne capturent pas correctement une large proportion de la variabilité du génome mitochondrial. Une approche ciblée, et non pangénomique était donc requise afin d'explorer dans quelle mesure la variabilité du génome mitochondrial influence le risque de cancer du sein. Dans ce contexte, j'ai montré que les porteurs de mutations pathogènes dans *BRCA2* appartenant à l'haplogroupe T1a1 ont 30% à 50% moins de risque de développer un cancer du sein que ceux appartenant à d'autres haplogroupes. Si cela est validé par d'autres études, cette nouvelle donnée représente une information significative dans le suivi clinique des femmes porteuses de mutation, et pourrait influencer la stratégie choisie afin de gérer le risque qu'elles développent un cancer du sein.

Trois raisons principales pourraient expliquer l'incapacité de l'algorithme utilisé à assigner un haplogroupe pour environ 9% des individus inclus dans notre étude. Tout d'abord, connaissant le taux de mutation spontanée élevé du génome mitochondrial, certaines des combinaisons de SNPs observées dans nos données ont pu apparaître relativement récemment dans la population générale. Les haplotypes correspondants peuvent ne pas être encore inclus dans l'arbre de référence PhyloTree. Deuxièmement, une seule erreur de génotypage pourrait conduire à observer des haplotypes chimériques qui n'existent pas en réalité, bien que la qualité très élevée de nos données de génotypage rendent cette hypothèse improbable. Enfin, l'arbre évolutif de référence PhyloTree est basé sur une reconstruction de la phylogénie par parcimonie. Cet arbre pourrait ne pas être optimal au niveau de certaines clades, en particulier celles pour lesquelles peu de séquences sont disponibles dans les bases de données publiques, comme c'est le cas pour les haplogroupes Africains[277]. En cas d'incertitude, le choix que nous avons fait d'assigner l'haplotype court considéré au plus récent ancêtre commun partagé par les haplogroupes présentant tous cet haplotype court nous permet d'améliorer la puissance statistique de nos analyses sans pour autant introduire de biais dans la démarche de détection d'association.

Afin d'estimer la précision de la méthode d'imputation des haplogroupes employée, nous l'avons appliquée sur un jeu de 630 séquences de génomes mitochondriaux d'origine européenne ou caucasienne dont l'haplogroupe est connu. L'imputation par notre méthode n'a pas abouti pour 6.7% des séquences issues de ce jeu de données (contre 9% dans nos données expérimentales). D'après l'observation de l'origine ethnique des 42 individus correspondants (Table 8), une proportion importante de ces séquences est associée à des individus appartenant à des

ethnies particulières telles que les Siciliens ou les Sardes, et dont la structure de la population et la variabilité génomique sont légèrement différentes de celles observées dans la population générale de leur pays d'origine et de celles de la population européenne[144].

Pour ce jeu de données test, notre méthode a été capable de retrouver l'haplogroupe référencé pour 83% des séquences testées. La précision atteint 100% pour les séquences appartenant aux haplogroupes I, J, K, T, U, W et X. Elle est de 70% pour les séquences appartenant à l'haplogroupe H, et tombe à 3% pour les séquences appartenant à l'haplogroupe V. Pour les 30% des séquences H et les 97 % des séquences pour lesquelles la méthode n'a pas retrouvé l'haplogroupe correct, l'haplogroupe imputé était HV ou HV7. Or ces deux haplogroupes sont des ancêtres communs aux clades H et V. Ainsi, avec les 92 SNPs à notre disposition, la méthode utilisée n'est pas capable de distinguer correctement les clades H et V.

Nous avons également analysé plus en détail la capacité de notre méthode à inférer correctement les sous-haplogroupes T1, T1a1 et T2. Les résultats sont également présentés dans la table 9. L'haplogroupe T1a1 a été imputé avec succès pour 18 des 23 séquences rattachées à l'haplogroupe T1. Les 5 autres séquences appartiennent aux haplogroupes T1a3 et T1b, deux sous-haplogroupes observés principalement en Mésopotamie, Anatolie, Égypte, Angleterre, Algérie, Grèce et Inde. Pour ces 5 séquences, l'haplogroupe T a été imputé. Il n'est pas étonnant que notre set de 92 SNPs ne réussisse pas à imputer des haplogroupes présents principalement au Moyen-Orient, ce set de SNPs étant plutôt dédié à l'identification d'haplogroupes européens. Enfin, sur le jeu de données test, 100 % des séquences des haplogroupes T1a1 et T2 ont été imputées avec succès.

Avec uniquement 129 loci génotypés sur les 16569 bases constituant le génome mitochondrial, il est certain que nous n'explorons pas entièrement la variabilité des haplotypes mitochondriaux. Une caractérisation plus précise du génome mitochondrial à l'échelle de l'individu nécessiterait d'utiliser des méthodes d'acquisition des données différentes, telles que le séquençage à haut débit - ou NGS pour *Next Generation Sequencing*. Cependant, les technologies NGS présentent elles aussi leurs propres limites et challenges. Ainsi, du fait de la séquence de certaines régions du génome, il est actuellement très difficile d'aligner certains fragments d'ADN séquencés sur le génome de référence, à cause par exemple de la forte homologie du génome mitochondrial avec le génome nucléaire. Un important traitement bioinformatique est nécessaire afin de surmonter ces contraintes technologiques. Enfin, même pour un génome de seulement 16 569 paires de bases comme celui de la mitochondrie, le séquençage de son intégralité pour plus de 20 000 individus représenterait une hausse majeure du coût de l'étude en comparison du génotypage de 129 SNPs.

L'association mise en évidence a été détectée parmi les porteuses de mutation sur *BRCA2*, mais n'a pas été observée chez les femmes mutées sur *BRCA1*. Cela pourrait révéler une altération des mécanismes liés spécifiquement à BRCA2 dans le développement du cancer. Ces altérations n'auraient pas ou peu d'impact pour les femmes présentant une protéine BRCA1 altérée et une protéine BRCA2 fonctionnelle, alors qu'elles modifierait significativement le risque dans le cas où BRCA2 est altérée. Il est aujourd'hui admis que les cancers associés à des mutations sur *BRCA1* et *BRCA2* sont différents du point de vue phénotypique. Les deux types de tumeurs correspondants ne présentent pas les mêmes profils d'expression génique ou les

mêmes altérations du nombre de copies de gènes[278]. BRCA1 et BRCA2 sont toutes deux impliquées dans la réparation des dommages à l'ADN via la recombinaison homologue des cassures double-brins, mais aussi des cassures simple-brin - ou SSB pour *Single Strand Break*, une type de dommage à l'ADN extrêmement fréquent pouvant être occasionné par l'exposition aux espèces oxygénées réactives[279]. Des études ont montré qu'en l'absence d'une protéine BRCA2 fonctionnelle, la réparation des cassures d'ADN par recombinaison homologue pouvait être compromise[280,281]. Récemment, une étude[282] a montré qu'un processus alternatif de réparation de l'ADN par recombinaison homologue - ou aHR pour *alternative Homologous Recombination*- était actif au niveau des cassures simple-brin, et que ce mécanisme était différent de celui observé au niveau des cassures double-brins. Comme un autre mécanisme de réparation appelé SSA - pour *Single Strand Annealing* - dont le principe est de joindre des portions de séquences répétées positionnées de part et d'autre de la cassure d'ADN) - ce mécanisme de réparation alternatif est dépendant des protéines RAD51 et BRCA2. Ces protéines exercent un rétrocontrôle négatif sur l'utilisation de ce mécanisme. La voie de réparation aHR est donc stimulée par l'inhibition de l'expression ou de l'activité de RAD51 et BRCA2. De plus, la voie aHR requiert la présence de la protéine BRCA1. Ainsi, la voie aHR peut être active dans le cas où le processus de recombinaison homologue classique est altéré ou inactif, ce qui peut être le cas chez les porteurs de mutations pathogènes sur *BRCA2*. De plus, la voie aHR favoriserait la perte d'hétérozygotie, qui contribue à l'apparition de nouvelles mutations et à la tumorogénèse. Ainsi, l'usage de la voie aHR pourrait expliquer pourquoi des pertes d'hétérozygotie sont plus fréquemment observées dans des tumeurs caractérisées par une déficience de la recombinaison homologue, ainsi que par des altérations de BRCA1/2 dans des tumeurs de l'ovaire ou dans des lignées cellulaires de cancer du sein et de la prostate[283]. L'association détectée concorderait donc avec l'hypothèse selon laquelle les individus T1a1 seraient moins exposés que les autres porteurs de mutations pathogènes sur *BRCA2* à l'instabilité génomique liée à l'utilisation de la voie aHR comme voie de réparation des cassures simple-brin. Cela concorderait également avec le fait que l'on ne retrouve pas cette association chez les individus mutés sur *BRCA1*.

Mueller et ses collaborateurs[284] ont étudié les différences fonctionnelles entre les haplogroupes T et H, l'haplogroupe le plus fréquent dans la population générale caucasienne. Ils ont généré des *cybrids*, en combinant des cellules issues de la lignée HEK293 (une lignée de cellules embryonnaires humaines de rein) dépourvue de leur ADN mitochondrial avec des thrombocytes de divers haplogroupes. Les thrombocytes, également appelés plaquettes, sont des cellules sanguines ne possédant pas de noyau, mais possédant des mitochondries dans leur cytoplasme. Les *cybrids* de l'haplogroupe T ont montré de meilleurs capacités de résistance au stress oxydatif et de survie suite à l'exposition au peroxyde d'hydrogène que les *cybrids* de l'haplogroupe H. Cependant, Amo[285]et ses collaborateurs n'ont mis en évidence aucune différence de synthèse ou de rendement bioénergétique au sein de la mitochondrie entre les *cybrids* des haplogroupes T et H possédant un génome nucléaire identique. Lin[19] et ses collaborateurs ont également utilisé des *cybrids* pour explorer la sensibilité des mitochondries de différents haplogroupes au stress oxydatif. Cependant, alors que l'étude de Mueller était focalisée sur les haplogroupes européens, l'étude de Lin a montré que l'haplogroupe asiatique B4b était plus sensible au stress oxydatif induit par une exposition au peroxyde d'hydrogène H_2O_2 que le reste des haplogroupes asiatiques. De plus, dans cette même étude, l'haplogroupe N9b a montré de meilleurs capacités de survie en situation de stress oxydatif. Cependant, aucun des variants constituant l'haplogroupe N9b ne semble coïncider avec ceux définissant l'haplogroupe T1a1. Néanmoins, cette étude fournit

des preuves fonctionnelles que les porteurs de l'haplogroupe T pourraient être moins sensibles au stress oxydatif et aux dommages à l'ADN associés, ce qui supporte l'observation que nous avons faite d'une association inverse entre l'haplogroupe T1a1 et le risque de cancer du sein chez les porteurs de mutations sur *BRCA2*.

Les polymorphismes T9899C, G11812A/rs41544217, et G13708A/rs28359178 ont été identifiés par ALTree comme trois sites de susceptibilité possibles potentiellement impliqués dans l'association détectée. G13708A est connu pour être une mutation secondaire de la neuropathie optique héréditaire de Leber, ou syndrome LHON. Bien que le rôle des mutations secondaires dans le syndrome LHON soit controversé, G13708A pourrait être associé à des altérations de la chaîne respiratoire mitochondriale dans le cadre de cette pathologie. De plus, G13708A a déjà été observé en tant que mutation somatique dans des tumeurs du sein, alors que ce polymorphisme était absent des tissus normaux adjacents et des échantillons sanguins[286]. Une grande proportion des mutations somatiques mitochondriales observées dans les tumeurs du sein sont également des polymorphismes mitochondriaux connus. Cela est cohérent avec l'hypothèse selon laquelle les cellules tumorales sont enclines à acquérir les mêmes mutations que celles ayant été acquises par sélection adaptative lorsque les premiers hommes ont migré hors d'Afrique et ont été confrontés à de nouveaux environnements[287]. Curieusement, le polymorphisme constitutionnel G13708A a déjà été montré inversement associé au risque de cancer du sein familial[225], avec un OR=0.47 (Intervalle de confiance à 95% : 0.24-0.92). L'OR que nous avons estimé à l'aide d'un modèle de Cox afin de quantifier l'association inverse entre l'haplogroupe T1a1 et le risque de cancer du sein est tout à fait comparable : OR=0.62 (95% CI=0.40-0.95).

L'ensemble de ces données suggère que l'haplogroupe T1a1 modifie le risque de cancer du sein. Des études approfondies des mécanismes biologiques sous-jacents à l'association détectée sont requises afin de renforcer l'hypothèse formulée selon laquelle le génome mitochondrial influence le risque de cancer du sein, et spécialement chez les femmes porteuses de mutation sur *BRCA2*. Si ce rôle est confirmé, alors la variabilité du génome mitochondrial pourrait devenir un facteur à prendre en compte dans le choix de la stratégie de gestion du risque de cancer du sein que ces femmes présentent.

Chapitre 3

Séquençage ciblé du génome mitochondrial de femmes avec antécédents familiaux pour le cancer du sein, mais non porteuses de mutations sur BRCA1/2

Le troisième axe de recherche auquel je me suis intéressée pendant ma thèse a consisté en la caractérisation par séquençage des variants constitutionnels du génome mitochondrial chez des femmes ayant des antécédents familiaux de cancer du sein, mais pour lesquelles aucune mutation pathogène sur les gènes *BRCA1* et *BRCA2* n'a été détectée.

Nous avons présenté auparavant le cancer du sein comme une maladie multifactorielle complexe, influencée aussi bien par des facteurs liés au style de vie que par des facteurs génétiques. La part génétique du risque total de cancer du sein est aujourd'hui estimée entre 5% et 7%[27]. Comme évoqué précédemment, *BRCA1* et *BRCA2* sont deux des gènes participant à la prédisposition génétique au cancer du sein. D'autres gènes et variants ont également été identifiés comme facteurs de prédisposition et comme modificateurs du risque initialement conféré par une mutation pathogène sur *BRCA1* et *BRCA2*. Les analyses de liaison génétique, les études de type gène candidat, ainsi que les études pangénomiques ont permis de mettre en évidence des variants génétiques expliquant environ 50% de l'héritabilité du cancer du sein. Cependant, les études les plus récentes ayant échoué à mettre en évidence de nouveaux gènes à forte pénétrance associés au risque de cancer du sein, il est de plus probable que la part encore non expliquée du risque de cancer du sein soit due à un effet cumulé de variants relativement fréquents à l'effet modéré ou faible, à des variants de pénétrance moyenne mais rares[288], et par l'effet synergique de certains variants génomiques conjugués avec l'exposition à des facteurs environnementaux ou liés au style de vie[289].

L'importance du stress oxidatif et du rôle potentiel de la mitochondrie, et donc des variants portés par son génome, dans l'apparition de l'instabilité génomique et dans le développement tumoral a été souligné dans les parties précédentes. Alors que les analyses de liaison génétique sont dédiées à l'étude de génomes diploïdes dans lesquels le processus de recombinaison est actif, la mitochondrie possède un génome haploïde. D'autre part, jusqu'à présent, les puces commerciales utilisées pour les études GWAS ne permettent pas de capturer la variabilité dans certaines partie du génome, notamment sur le génome mitochondrial. D'autres méthodes, telles que le séquençage ciblé mis en place dans cette étude sont donc requises afin d'explorer dans quelle mesure la variabilité du génome mitochondrial influe sur le risque de cancer du sein.

Dans cette optique, l'étude présentée ci-après a été mise en place afin de caractériser la variabilité du génome mitochondrial observée chez des individus diagnostiqués pour un cancer du sein et présentant des antécédents familiaux, mais chez lesquels aucune mutation pathogène n'a été détectée sur *BRCA1* et *BRCA2*.

I. Étude principale

I.1 Matériel

I.1 .1 L'étude GENESIS

L'étude GENESIS est une cohorte nationale de femmes françaises, coordonnée par Dominique Stoppa-Lyonnet et Nadine Andrieu à l'institut Curie à Paris. Cette étude a été mise en place afin d'identifier de nouveaux gènes de susceptibilité au cancer du sein. L'étude GENESIS inclut des paires de soeurs diagnostiquées pour un cancer du sein, mais testées négatives pour les mutations pathogènes sur *BRCA1/2*. Elle inclut également des contrôles appariés non apparentés partageant le même environnement, ou des amies proches du cas index inclus.

I.1 .2 Sélection des échantillons

Les échantillons séquencés ont été sélectionnés parmi les cas index non apparentés inclus dans l'étude. Les cas index ont été classés en fonction de l'agressivité de leur phénotype, en particulier en fonction le l'âge au diagnostic et de la bilatéralité de leur cancer, c'est à dire de l'atteinte d'un seul ou des deux seins. Ces caractéristiques sont résumées pour les individus séquencés dans les Tables 14 et 15. Des échantillons sanguins ont été prélevés, et sont centralisés au laboratoire de Génétique du cancer du sein au Centre de recherche en Cancérologie de Lyon (CRCL), où les ADNs ont été extraits.

Table 14 – Caractéristiques des individus séquencés

Âge au 1er cancer du sein du cas index inclus *	46.1 ± 7.9
Âge au 1er cancer du sein de la soeur incluse*	45.7 ± 6.7
Bilatéralité du cancer pour les cas index	Oui : 51.2% Non : 48.8%

* en années, moyenne ± écart-type

I.1 .3 Séquençage du génome mitochondrial

La préparation des librairies et le séquençage ont été effectués par des collaborateurs sur la plateforme de séquençage du Centre Léon Bérard. Des amorces correspondant à 11 amplicons ont été conçues afin de couvrir l'intégralité du génome mitochondrial. Leur spécificité contre le génome mitochondrial a été testée, et aucune n'amplifie une région de forte homologie avec le génome nucléaire. Les librairies de séquençage ont été préparées à l'aide des kits *NEBNext Fast DNA Fragmentation & Library Prep Set* pour Ion Torrent. Les échantillons d'ADN ont été fragmentés, la taillé ciblée étant 175 paires de bases. Les échantillons ont été multiplexés par 48 : pour chaque run de séquençage, un code barre spécifique à chacun des 48 échantillons séquencés dans un run est accolé aux fragments d'ADN de l'échantillon correspondant. Les librairies ont été chargées sur des puces 316 et séquencées sur un séquenceur Ion Torrent *Personalized Genome Machine*, ou PGM. Un run de séquençage correspond au séquençage des échantillons déposés sur une puce. Une puce ayant une capacité limitée, 10 runs de séquençage ont été nécessaires pour séquencer les 436 échantillons sélectionnés.

Table 15 – Catégories d'inclusion des cas index séquencés

Catégorie d'inclusion	**Effectifs**
Cancer du sein bilatéral du cas index , diagnostiqué avant 50 ans Cancer du sein bilatéral de la soeur, diagnostiqué avant 50 ans	8
Cancer du sein bilatéral du cas index , diagnostiqué avant 60 ans Cancer du sein bilatéral de la soeur	8
Cancer du sein du cas index diagnostiqué avant 60 ans Cancer du sein bilatéral de la soeur	36
Cancer du sein bilatéral du cas index , diagnostiqué avant 50 ans Cancer du sein unilatéral de la soeur, diagnostiqué avant 50 ans	23
Cancer du sein unilatéral du cas index, diagnostiqué avant 50 ans Cancer du sein bilatéral de la soeur, diagnostiqué avant 50 ans	23
Cancer du sein unilatéral du cas index, diagnostiqué avant 50 ans Cancer du sein unilatéral de la soeur, diagnostiqué avant 50 ans	164
Cancer du sein bilatéral du cas index , diagnostiqué avant 60 ans	174
Total	**436**

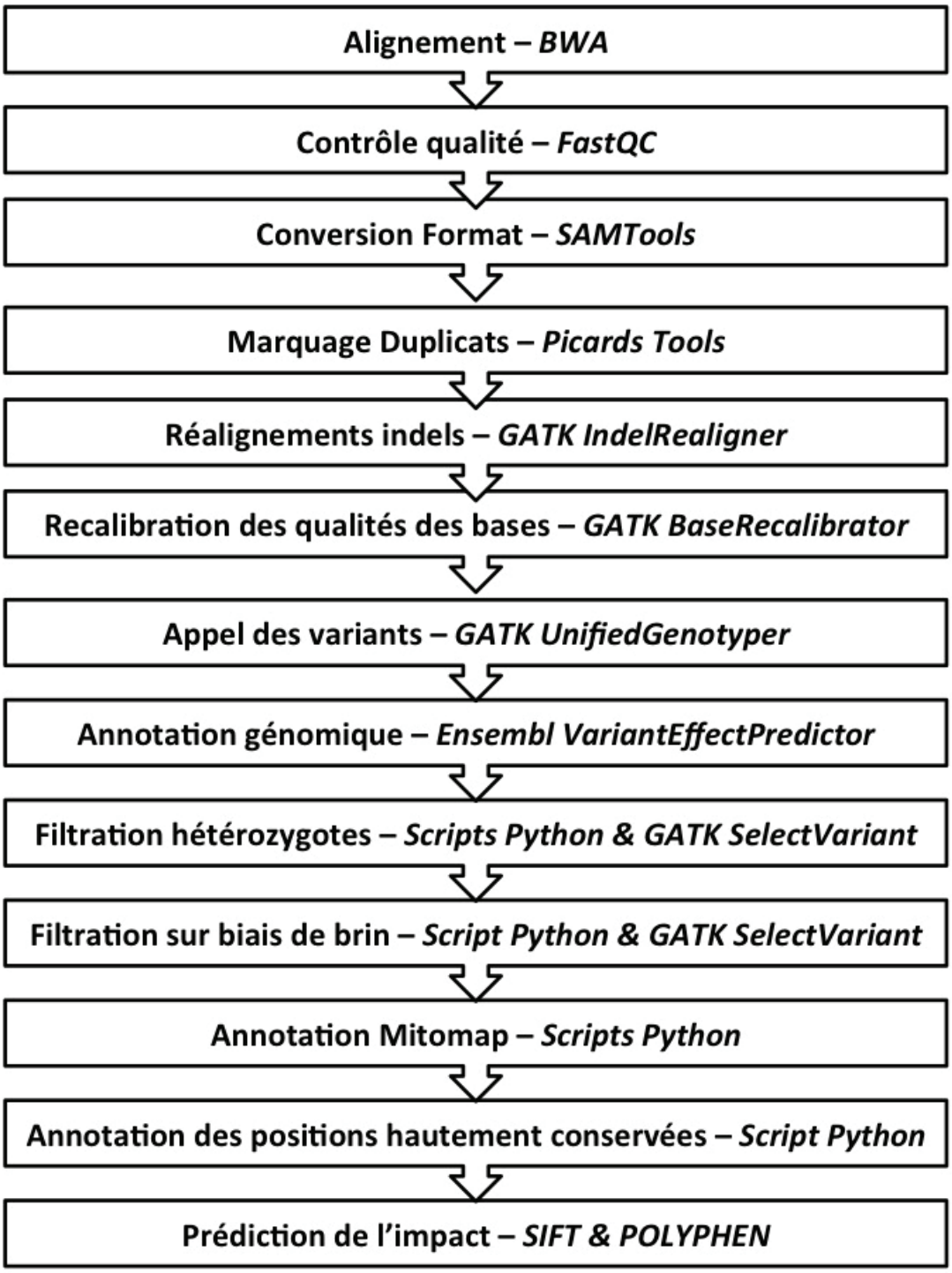
I.2 Méthodes

I.2 .1 Développement et automatisation du pipeline d'analyses bioinformatiques

J'ai mis un place un pipeline automatique afin d'analyser les données générées par chaque run de séquençage. Les fichiers récupérés en sortie du séquenceur contiennent les séquences nucléotidiques des fragments d'ADN séquencés - également appelés lectures ou *reads* - ceux-ci étant regroupés par échantillon. L'ensemble des analyses s'effectue sur l'intégralité des reads de chaque échantillon. L'intégralité du pipeline d'analyses mis en place est schématisé sur la figure 38. Ce pipeline d'analyse s'exécute consécutivement sur chacun des échantillons multiplexés dans un même run de séquençage. Lorsque l'intégralité des échantillons d'un run a été analysée, un rapport automatique est généré au format PDF. De même, à l'issue de l'analyse de chaque run, la table complète des variants détectés dans l'ensemble de l'étude sur les données de séquençage est mise à jour.

Tout d'abord, les reads sont alignés contre la séquence de référence du génome mitochondrial rCRS déposée sur GenBank sous le numéro d'accession NC_012920.1 avec l'algorithme d'alignement BWA[290] (v0.7.5a). Une étude complémentaire a d'autre part été réalisée afin de déterminer quel algorithme d'alignement était le plus approprié dans le cadre de cette étude, les résultats sont présentés en section II. du chapitre courant. La qualité des reads a été analysée à l'aide de la suite d'outils FastQC[291](v0.10.1). Les outils Samtools[292] (v0.1.19), Picard[293]

Figure 38 – Pipeline d'analyses mis en place détaillant chaque étape ainsi que l'outil utilisé pour la réaliser



(v1.96) et GATK[294,295] (v2.5-2) ont été utilisés pour le traitement des données après alignement. Un premier affinage de l'alignement a été effectué avec l'outil GATK *IndelRealigner* en réalignant localement les reads positionnés au niveau des insertions et des délétions détectées à l'issue de l'alignement initial. Le score de qualité de chaque base séquencée a été recalibré avec l'outil GATK *BaseRecalibrator*, un outil permettant d'affiner les scores de qualité en prenant en compte divers facteurs relatifs à certains paramètres du séquençage et au contexte nucléotidique avoisinant. L'appel des variants par rapport au génome de référence, c'est à dire l'action de déterminer l'ensemble des positions du génome qui d'après les données de séquençage, présentent des variations par rapport au génome de référence, a été effectué avec l'algorithme GATK *UnifiedGenotyper*. Les variants détectés ont ensuite été annotés et filtrés, d'une part avec l'outil *Variant Effect Predictor* d'Ensembl[296](projet international dédié à l'annotation automatique du génome humain et d'autres génome eucaryotes), et d'autre part à l'aide de scripts python développés personnellement.

I.2 .2 Annotation et filtration des variants détectés

La méthodologie générale employée afin de détecter des variants génomiques en comparaison d'un génome de référence est décrite ci-après. Comme présenté au paragraphe précédent, après l'alignement initial et un potentiel réalignement dans certaines zones plus complexes telles que celles où des insertions/délétions ont été détectées en premier lieu, on applique un algorithme qui effectue l'appel des variants. Il existe plusieurs algorithmes d'appel des variants qui reposent sur différentes approches : les modèles mis en place par ces algorithmes sont complexes, et ne seront pas détaillés ici. Cependant, le principe qui les régit est le même : si un variant est appelé par l'algorithme, alors c'est que d'après les données de séquençage, la probabilité que ce variant soit réel est suffisamment élevée pour le considérer comme un variant potentiel. Ainsi, la plupart des algorithmes d'appel de variants a tendance à être très sensible, mais peu spécifique. Ces algorithmes fournissent donc généralement une liste de variants potentiels très longue, dont seulement une faible proportion existe vraiment.

Plusieurs étapes de filtration ont donc été mises en place afin d'éliminer les probables faux-positifs. Tout d'abord, la mitochondrie ayant un génome haploïde, les variants hétérozygotes sont jugés peu fiables et filtrés. D'autre part, dans certains cas, l'existence d'un variant n'est attestée que par des lectures s'alignant dans un seul sens sur le génome de référence ou pour lesquelles la proportion de lectures s'alignant dans les deux sens est déséquilibrée. On appelle cela le biais de brin. Or, ces variants sont souvent des artefacts dus à des erreurs dans l'alignement local, ils sont donc exclus.

On annote les variants restants selon qu'ils sont déjà connus ou non dans la base de données MITOMAP. Un variant inconnu jusqu'ici peut être indicateur de plusieurs choses : un variant réel, bien que non-référencé, ce qui est vraisemblable s'il est peu fréquemment observé, si c'est un SNP rare par exemple. Dans le cas où le nouveau variant est observé fréquemment dans l'étude, ou si ce variant est une caractéristique spécifique de la pathologie étudiée, ou (ce qui est plus vraisemblable) si ce variant résulte d'une erreur de séquençage/d'alignement systématique. Ainsi, savoir si un variant est connu et déjà référencé dans un contexte précis peut permettre de nuancer sa vraisemblance.

On annote aussi les variants selon qu'ils sont ou non localisés sur des positions du génome mitochondrial fortement conservées au cours de l'Évolution. La détermination de ces positions a fait l'objet d'une seconde étude complémentaire présentée en partie III. du chapitre courant. Il est utile de savoir si un variant affecte une position fortement conservée au cours de l'Évolution. En effet, on s'attend à ce des mutations apparaissent au sein du génome mitochondrial, d'autant plus que celui-ci est caractérisé par un taux de mutation élevé. Si une position est fortement conservée, c'est qu'une pression de sélection purifiante s'exerce à cet endroit, et que les changements apparaissant sont contre-sélectionnés. La séquence nucléotidique à cette position code sans doute pour des éléments structuraux ou régulateurs, protéiques ou ribonucléiques, ayant un rôle important. Il est donc d'autant moins probable d'observer des polymorphismes à cette position. Cependant, si les données sont robustes et que la mutation ou le polymorphisme semble être réel, alors les effets induits par ce variant peuvent être très dommageables.

Enfin, pour les variants détectés sur des séquences codantes, leur effet au niveau du transcrit ou de la protéine est prédit à l'aide de deux algorithmes distincts, SIFT[109] et PolyPhen[110]. Ces deux algorithmes de prédiction se basent sur des informations telles que la localisation précise de la mutation, ses éventuelles répercussions directes sur la protéine comme l'induction d'un décalage de phase, la modification de la structure 3D de la protéine, ou encore l'atteinte d'un site actif, pour estimer la sévérité des conséquences induites par cette mutation. L'effet prédit par SIFT peut être « délétère » ou « bien toléré », alors que l'effet prédit par PolyPhen peut être « bénin », « potentiellement dommageable », ou bien « probablement dommageable ».

I.2 .3 Analyses

Une analyse du taux d'enrichissement global en variants par gène a été effectuée. Cette analyse consiste à calculer le rapport entre le nombre de variants distincts détectés sur un gène donné, rapporté à la longueur totale du gène. Soit :

- N_x le nombre de variants distincts détectés sur un gène donné x
- l_x la longueur du gène x en paires de bases

Alors pour un gène x donné, le taux d'enrichissement global r_g en variants distincts par Mb se calcule de la manière suivante :

$$r_g = \frac{N_x}{l_x} \times 1000$$

Ce calcul a été effectué pour tous les gènes portés par l'ADN mitochondrial qui ne codent pas pour les ARNs de transfert de la mitochondrie ; ces gènes étant extrêmement courts, le rapport obtenu aurait été biaisé de manière importante.

Selon le même principe, le taux d'enrichissement en variants par gène pondéré à été calculé. En effet, la fréquence à laquelle chaque variant est observé dans notre étude n'entre pas en compte dans le calcul du taux d'enrichissement global en variants distincts. Ainsi, qu'un variant ait été observé 1 ou 100 fois parmi les 436 échantillons analysés ne change en rien le résultat du taux d'enrichissement global calculé pour le gène considéré. C'est pourquoi le taux d'enrichissement pondéré a été introduit.

Soit :

- N_x le nombre de variants distincts détectés sur un gène donné x
- i l'indice de chacun des N_x variants détectés sur le gène x
- n_i le nombre d'occurrences observées du variant i parmi les 436 échantillons GENESIS analysés
- l_x la longueur du gène x en paires de bases

Alors pour un gène x donné, le taux d'enrichissement pondéré r_w en variants par Mb se calcule de la manière suivante :

$$r_w = \frac{\sum_{i=1}^{N_x} n_i}{l_x} \times 1000$$

Ainsi, le taux d'enrichissement pondéré r_w tient compte à la fois de la variabilité observée sur l'ensemble du gène et de la fréquence de chaque variant observé.

I.3 Résultats

La Table 16 présente des statistiques effectuées sur les données post-séquençage. Au total, au cours de cette étude, environ 20 millions de reads ont été séquencés, d'une longueur moyenne de 137 bases. En moyenne, 47 000 reads environ ont été séquencés pour chaque échantillon, sur lesquels un peu plus de 27 000 ont pu être alignés contre le génome de référence, ce qui correspond à un taux d'alignement moyen de 57.4%. Ce taux d'alignement est relativement faible, mais l'étude complémentaire présentée en section II. a indiqué qu'il est plus judicieux d'utiliser BWA plutôt que TMAP pour effectuer l'alignement dans le cadre de notre étude. BWA est certes plus stringent que TMAP, mais cette stringence garantit un alignement correct des reads, alors que TMAP est plus permissif. Or un alignement plus fiable permet d'effectuer l'appel des variants avec moins d'erreurs. On préfère donc avoir moins de reads alignés, mais que l'alignement de ces reads soit de bonne qualité.

Table 16 – Statistiques post-séquençage

Nombre total de reads	20 514 117
Nombre de reads séquencés par échantillon*	47 051 ± 16 657
Nombre de reads alignés par échantillon*	27 453 ± 10 704
Longueur des reads séquencés**	137 ± 55

* moyenne ± écart-type

** en paires de bases, moyenne ± écart-type

La Table 17 présente la proportion du génome mitochondrial couverte à différentes profondeurs, en moyenne sur l'ensemble des échantillons. Ainsi, moins de 1% du génome mitochondrial est couvert avec une profondeur inférieure à 5X. C'est à dire que pour 99% des positions constituant le génome mitochondrial, plus de 5 reads s'alignent sur celles-ci localement. De même, 10.7 % du génome mitochondrial est couvert avec une profondeur inférieure à 50X, ce qui signifie que pour environ 90% des positions constituant le génome mitochondrial, plus de 50 reads s'alignent sur ces positions. Une profondeur de 50 X est tout à fait correcte pour effectuer l'appel des variants, d'autant plus sur un génome haploïde et lorsqu'on fait abstraction de l'hétéroplasmie, situation dans laquelle on ne cherche pas à détecter les variants hétérozygotes. La couverture observée sur le génome mitochondrial est donc suffisante, malgré le faible taux d'alignement observé.

Table 17 – Couverture du génome mitochondrial après séquençage

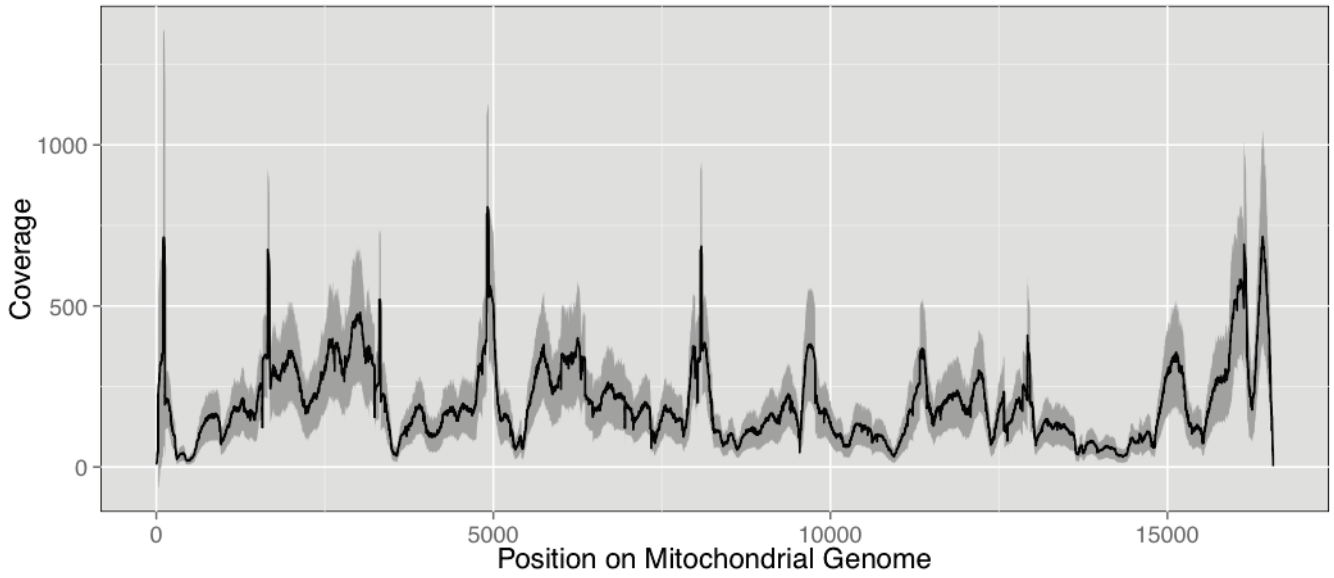
Seuil de couverture T	**Proportion du génome mitochondrial ayant une couverture** $< T$ *
5 X	0.3 ± 1.9
10 X	0.8 ± 3.7
20 X	2.4 ± 7.0
50 X	10.7 ± 13.3

* en %, moyenne ± écart-type

La Figure 39 représente la distribution de la couverture le long du génome mitochondrial pour l'ensemble des échantillons. La couverture observée est très variable. Certaines positions sont couvertes à plus de 1 000 X, alors que d'autres sont très faiblement couvertes.

La Table 18 détaille le nombre et le type des variants qui ont été détectés sur l'ensemble des 436 échantillons GENESIS séquencés. 1 157 variants distincts passent les critères de qualité des filtres mis en place dans le pipeline décrit plus haut. Plus de 99% des variants détectés sont des substitutions. Une très grande proportion des 1 157 variants sont déjà connus, mais 3% ne sont pas référencés dans MITOMAP. Plus de 80% des variants détectés sont situés sur des gènes. Cette tendance est contraire à celle observée généralement sur le génome nucléaire, sur lequel la proportion des variants géniques ne dépasse pas les 10% en règle générale, mais le génome mitochondrial est une structure à part : les gènes qu'il supporte sont organisés de manière extrêmement dense. Le génome mitochondrial possède peu de séquences intergéniques, ce qui n'est pas sans rappeler sa probable origine bactérienne.

Figure 39 – Couverture après alignement le long du génome mitochondrial



Représentation de la couverture moyenne par échantillon le long du génome mitochondrial. *Courbe noire* : couverture moyenne calculée sur les 436 échantillons. *Surface grisée* : Moyenne ± écart-type de la couverture. La couverture des positions situées sur les régions de chevauchement de deux amplicons successifs a été divisée par 2.

Table 18 – Description globale des variants détectés

Total	1157	*100* %
Substitutions	1147	*99.1* %
Insertions	5	*0.4* %
Délétions	5	*0.4* %
Variants répertoriés sur MITOMAP	1122	*97.0* %
Variants non répertoriés sur MITOMAP	35	*3.0* %
Variants intergéniques	213	*18.4* %
Variants géniques	944	*81.6* %
Effet des substitutions situés sur des gènes		
Synonymes	542	*46.8* %
Faux sens	240	*20.7* %
Non codantes	165	*14.3* %
Perte d'un codon stop	1	*0.1* %

La Table 19 est une table de contingence des variants localisés sur des régions codant pour des protéines, en fonction de leur impact sur la chaîne d'acides aminés et la protéine globale, prédit d'après les deux algorithmes de prédiction appliqués SIFT et PolyPhen. À noter que 24 variants sont prédits « délétères » par SIFT et « probablement dommageables » par PolyPhen. Ces 24 variants sont décrits en détail dans la Table 20.

Table 19 – Prédiction de l'effet des substitutions géniques par SIFT et PolyPhen

		SIFT	
		Toléré	Délétère
PolyPhen	Bénin	143	39
	Potentiellement dommageable	4	11
	Probablement dommageable	11	24

La Table 21 détaille les variants qui ne sont pas référencés dans MITOMAP. Aucun de ces variants n'a été détecté avec une haute fréquence parmi les 436 échantillons séquencés, ce qui laisse supposer que ces variants ne sont pas des artefacts des données de séquençage. La plupart de ceux-ci n'affectent pas des positions conservées au cours de l'Évolution, ce qui est également compatible avec le fait que ce soit des polymorphismes rares.

La Table 22 présente le taux d'enrichissement global en variants distincts par gène r_g, ainsi que le taux d'enrichissement en variants par gène pondéré r_w, la calcul de ces indices étant explicité dans la section Méthodes. La moyenne du taux d'enrichissement global r_g est de 64.2 variants/Mb. La moyenne du taux d'enrichissement pondéré r_w est de 0.44 variants.individu/Mb. Les gènes *MT-CYB* et *MT-ATP6* présentent les valeurs du taux d'enrichissement global les plus élevées, 84.1 et 86.5 variants/Mb respectivement. Les valeurs les plus faibles de r_g sont 35.6 et 37.2 variants/Mb, et sont respectivement observées pour les gènes *MT-RNR1* et *MT-RNR2*. Concernant le taux d'enrichissement pondéré, la valeur la plus élevée observée est 1.07 pour le gène *MT-RNR1*, suivies par les valeurs 0.92 et 0.91 pour les gènes *MT-ATP6* et *MT-CYB*. La valeur la plus faible observée est 0.11 pour le gène *MT-CO2*. La comparaison de ces deux indices peut permettre de nuancer l'enrichissement en variants de certains gènes. Par exemple, les gènes *MT-CO2* et *MT-ND2* ont tous les deux des taux d'enrichissement global comparables, ayant respectivement pour valeur 61.4 et 64.2 variants/Mb. Cependant, leurs taux d'enrichissement pondéré valent respectivement 0.62 et 0.11. En somme, bien que rapporté à la longueur du gène, un nombre comparable de variants ait été détecté sur chacun de ces deux gènes, beaucoup plus d'individus portent un variant sur le gène *MT-CO2* que sur le gène *MT-ND2*.

Table 20 – Description des variants prédits « délétères » par SIFT et « probablement dommageables » par PolyPhen

Pos.	Ref.	Alt.	rsID	MITOMAP	Nb	Freq.	Conservé	Gène	Effet	Chang. de codon	Chang. d'a.a.
3388	C	A	.	Connu	1	0.0023	✗	MT-ND1	missense	Cta/Ata	L/M
6237	C	A	.	Connu	2	0.0046	✗	MT-CO1	missense	Ctg/Atg	L/M
6489	C	A	rs28461189	Connu	3	0.0069	✓	MT-CO1	missense	Ctc/Atc	L/I
7941	A	G	.	Connu	1	0.0023	✓	MT-CO2	missense	aAc/aGc	N/S
7964	T	C	.	Connu	1	0.0023	✗	MT-CO2	missense	Ttc/Ctc	F/L
7976	G	A	.	Connu	1	0.0023	✓	MT-CO2	missense	Ggc/Agc	G/S
8563	A	G	.	Connu	1	0.0023	✗	MT-ATP6	missense	Aca/Gca	T/A
8839	G	A	.	Connu	2	0.0046	✓	MT-ATP6	missense	Gcc/Acc	A/T
8920	G	A	.	Connu	1	0.0023	✗	MT-ATP6	missense	Ggc/Agc	G/S
9010	G	A	.	Connu	1	0.0023	✓	MT-ATP6	missense	Gct/Act	A/T
9448	A	G	.	Connu	1	0.0023	✓	MT-CO3	missense	tAc/tGc	Y/C
9500	C	A	.	Inconnu	1	0.0023	✗	MT-CO3	missense	ttC/ttA	F/L
9577	T	C	.	Inconnu	1	0.0023	✓	MT-CO3	missense	cTa/cCa	L/P
9903	T	C	rs199999390	Connu	1	0.0023	✓	MT-CO3	missense	Ttt/Ctt	F/L
11087	T	C	.	Connu	1	0.0023	✓	MT-ND4	missense	Ttc/Ctc	F/L
12634	A	G	.	Connu	3	0.0069	✓	MT-ND5	missense	Atc/Gtc	I/V
12923	G	T	.	Connu	1	0.0023	✗	MT-ND5	missense	tGa/tTa	W/L
13129	C	T	.	Connu	1	0.0023	✓	MT-ND5	missense	Ccc/Tcc	P/S
13973	A	T	.	Connu	1	0.0023	✗	MT-ND5	missense	cAa/cTa	Q/L
14180	T	C	.	Connu	2	0.0046	✗	MT-ND6	missense	tAt/tGt	Y/C
14484	T	C	.	Connu	2	0.0046	✗	MT-ND6	missense	Atg/Gtg	M/V
14769	A	G	rs28357679	Connu	2	0.0046	✓	MT-CYB	missense	aAc/aGc	N/S
15218	A	G	rs2853506	Connu	10	0.0229	✗	MT-CYB	missense	Aca/Gca	T/A
15773	G	A	.	Connu	1	0.0023	✓	MT-CYB	missense	Gta/Ata	V/M

Pos. : Position sur le génome mitochondrial
Ref. : Allèle de référence
Alt. : Allèle alernatif
rsID : Identifiant dbSNP s'il existe
MITOMAP : ce variant est-il répertorié dans MITOMAP ?
Nb : Nombre d'individus porteurs de ce variant
Freq. : Fréquence de ce variant dans l'étude
Conservé : ce variant affecte-t-il une position strictement conservée chez les eucaryotes au cours de l'Évolution ?
Gène : Gène sur lequel se situe le variant
Effet : Effet du variant
Chang. de codon : Changement de codon provoqué par le variant
Chang. d'a.a. : Changement d'acide aminé provoqué par le variant

Table 21 – Description des variants non répertoriés dans MITOMAP

Pos.	Ref.	Alt.	rsID	Conservé	Effectif	Gène	Effet	SIFT	PolyPhen
393	T	A	.	✗	1	-	-	-	-
1713	A	G	.	✗	1	MT-RNR2	Exon non codant	-	-
1807	T	C	.	✗	1	MT-RNR2	Exon non codant	-	-
2150	T	TA	.	✓	1	MT-RNR2	Exon non codant	-	-
2156	A	AT	.	✗	1	MT-RNR2	Exon non codant	-	-
3385	A	T	.	✗	2	MT-ND1	faux sens	délétère	potentiellement dommageable
4875	C	T	.	✗	1	MT-ND2	synonyme	-	-
5573	A	G	.	✗	1	MT-TW	Exon non codant	-	-
5746	GA	G	.	✗	1	-	-	-	-
6113	A	T	.	✗	1	MT-CO1	synonyme	-	-
6200	C	T	.	✗	1	MT-CO1	synonyme	-	-
6569	C	T	.	✗	1	MT-CO1	synonyme	-	-
6608	C	T	.	✗	1	MT-CO1	synonyme	-	-
6812	A	G	.	✗	1	MT-CO1	synonyme	-	-
7004	A	G	.	✗	1	MT-CO1	synonyme	-	-
7366	C	T	.	✓	1	MT-CO1	faux sens	toléré	bénin
8263	C	T	.	✗	1	MT-CO2	synonyme	-	-
8465	C	T	.	✗	1	MT-ATP8	faux sens	toléré	probablement dommageable
8673	A	G	.	✗	1	MT-ATP6	synonyme	-	-
9138	C	T	.	✗	1	MT-ATP6	synonyme	-	-
9370	A	T	.	✗	1	MT-CO3	faux sens	toléré	bénin
9500	C	A	.	✗	1	MT-CO3	faux sens	délétère	probablement dommageable
9577	T	C	.	✓	1	MT-CO3	faux sens	délétère	probablement dommageable
9873	C	A	.	✗	1	MT-CO3	faux sens	toléré	probablement dommageable
9890	A	G	.	✗	1	MT-CO3	synonyme	-	-
10030	C	T	.	✗	1	MT-TG	Exon non codant	-	-
10094	C	A	.	✗	1	MT-ND3	synonyme	-	-
12098	C	T	.	✗	1	MT-ND4	synonyme	-	-
12266	A	G	.	✗	1	MT-TL2	Exon non codant	-	-
13380	C	T	.	✗	1	MT-ND5	synonyme	-	-
13792	C	T	.	✗	1	MT-ND5	synonyme	-	-
13806	C	T	.	✗	1	MT-ND5	synonyme	-	-
15620	C	T	.	✓	1	MT-CYB	faux sens	délétère	potentiellement dommageable
16229	T	A	.	✗	1	-	-	-	-
16454	C	T	.	✗	1	-	-	-	-

Pos. : Position sur le génome mitochondrial
Ref. : Allèle de référence
Alt. : Allèle alernatif
rsID : Identifiant dbSNP s'il existe
Conservé : ce variant affecte-t-il une position strictement conservée chez les eucaryotes au cours de l'Évolution ?
Gène : Gène sur lequel se situe le variant
Effet : Effet du variant
SIFT : Prédiction de l'impact du variant par SIFT
PolyPhen : Prédiction de l'impact du variant par PolyPhen

Table 22 – Distribution des variants détectés et enrichissement global en variants distincts par gène mitochondrial

Gène	Coordonnées	Type	Longueur (en pb)	Nb Variants	r_g	r_w
MT-TF	574-645	ARNt-mt	72	3	-	-
MT-RNR1	645-1599	ARNr-mt	955	34	35.6	1.07
MT-TV	1599-1668	ARNt-mt	70	1	-	-
MT-RNR2	1668-3228	ARNr-mt	1561	57	37.2	0.64
MT-TL1	3228-3303	ARNt-mt	76	1	-	-
MT-ND1	3305-4261	Protéique	957	58	60.6	0.27
MT-TI	4261-4330	ARNt-mt	70	3	-	-
MT-TQ	4327-4399	ARNt-mt	73	3	-	-
MT-TM	4400-4468	ARNt-mt	69	2	-	-
MT-ND2	4468-5510	Protéique	1043	64	61.4	0.62
MT-TW	5510-5578	ARNt-mt	69	5	-	-
MT-TA	5585-5654	ARNt-mt	70	6	-	-
MT-TN	5655-5728	ARNt-mt	74	0	-	-
MT-TC	5759-5825	ARNt-mt	67	5	-	-
MT-TY	5824-5890	ARNt-mt	67	0	-	-
MT-CO1	5902-7444	Protéique	1543	98	63.5	0.29
MT-TS1	7444-7513	ARNt-mt	70	2	-	-
MT-TD	7516-7584	ARNt-mt	69	4	-	-
MT-CO2	7584-8268	Protéique	685	44	64.2	0.11
MT-TK	8293-8363	ARNt-mt	71	3	-	-
MT-ATP8	8364-8571	Protéique	208	15	72.1	0.12
MT-ATP6	8525-9206	Protéique	682	59	86.5	0.92
MT-CO3	9205-9989	Protéique	785	60	76.4	0.23
MT-TG	9989-10057	ARNt-mt	69	7	-	-
MT-ND3	10057-10403	Protéique	347	24	69.2	0.30
MT-TR	10403-10468	ARNt-mt	66	3	-	-
MT-ND4L	10468-10765	Protéique	298	20	67.1	0.23
MT-ND4	10758-12136	Protéique	1379	85	61.6	0.46
MT-TH	12136-12205	ARNt-mt	70	3	-	-
MT-TS2	12205-12264	ARNt-mt	60	1	-	-
MT-TL2	12264-12335	ARNt-mt	72	2	-	-
MT-ND5	12335-14147	Protéique	1813	134	73.9	0.27
MT-ND6	14147-14672	Protéique	526	26	49.4	0.24
MT-TE	14672-14741	ARNt-mt	70	4	-	-
MT-CYB	14745-15886	Protéique	1142	96	84.1	0.91
MT-TT	15886-15952	ARNt-mt	67	14	-	-
MT-TP	15954-16022	ARNt-mt	69	2	-	-

r_g : Enrichissement global (en variants/Mb)

r_w : Enrichissement pondéré (en variants.individu/Mb)

I.4 Discussion

Dans cette étude, j'ai effectué une caractérisation précise de la variabilité du génome mitochondrial chez 436 femmes diagnostiquées pour un cancer du sein, ayant une histoire familiale de cancer du sein, mais testées négatives pour les mutations pathogènes identifiées sur *BRCA1* et *BRCA2*. Ces femmes ont été sélectionnées d'après la précocité et la sévérité de leur cancer.

Le séquençage du génome mitochondrial des femmes inclues dans l'étude a été réalisé avec la technologie Ion Torrent. Bien que le profil de couverture ne soit pas homogène le long du génome mitochondrial, ce profil est robuste dans les 10 runs de séquençage indépendants effectués dans cette étude. De plus, nous avons une profondeur moyenne supérieure à 50 X pour plus de 90% du génome mitochondrial. Ce profil de couverture ne peut pas être expliqué par les bornes des amplicons utilisés, potentiellement amplifiés de manière différente les uns des autres, ce qui pourrait conduire à des différences de couverture sur des segments du génome mitochondrial. Le taux de GC local ne permet pas non plus d'expliquer ce profil de couverture, tout du moins de manière linéaire. En effet, un test de corrélation de Pearson a été effectué entre la couverture observée et le taux de GC local calculé sur une fenêtre glissante de 100 paires de bases par pas de 50 paires de bases le long du génome mitochondrial. On observe une corrélation significative (p-value = $1.1.10^{-13}$), mais avec un coefficient de corrélation faible : $r = 0.058$, intervalle de confiance à 95% : [0.043 - 0.073]. La corrélation observée explique donc $r^2 < 0.003$, soit moins de 1% de la variabilité des données observées. Un modèle de régression quadratique n'a pas donné de meilleurs résultats. De plus, le génome mitochondrial présente peu de régions répétées, régions sur lesquelles les reads ne s'alignent généralement pas ou mal. Ce facteur ne peut pas non plus être responsable du profil de couverture observé.

Un article datant de Décembre 2011[297], accessible par *ce lien*, et s'intitulant *Ion Torrent Sequencing on Humans* a été publié sur le blog consacré à la Bioinformatique *Biolektures*. Cet article commente la publication par *Life Technologies* - la firme propriétaire de la technologie Ion Torrent - des données de séquençage humain issues de 2 runs tests (C18-99 et C24-141). Ce séquençage a été effectué sur une puce 318 à partir de l'ADN constitutionnel issu d'un échantillon sanguin d'un individu sain. L'auteur commente les résultats obtenus, en se focalisant sur le génome mitochondrial, et illustre notamment son article à l'aide du profil de couverture obtenu après séquençage et alignement avec BWA. Afin de mieux visualiser leurs similitudes et leurs différences, les profils de couverture obtenus pour chacun des 2 runs de séquençage et le profil de couverture global obtenu dans notre étude ont été juxtaposés, le résultat est présenté en Figure 40.

Plusieurs commentaires se dégagent de l'observation de la figure 40. Tout d'abord, la variabilité observée entre les courbes bleue et rouge est faible, ce qui signifie que le profil de couverture obtenu est robuste entre plusieurs runs de séquençage effectués dans les mêmes conditions, ce que nous observons également dans notre étude entre les 10 runs effectués. D'autre part, bien que les variations d'amplitude ne soient pas exactement similaires, les profils de couverture obtenus pour les deux expériences de séquençage se ressemblent énormément : les régions faiblement couvertes sont localisées au même endroit, et les courbes suivent quasiment les mêmes variations. Les kits utilisés pour la fragmentation de l'ADN et la préparation des librairies ne sont pas les mêmes. De plus, le séquençage du génome mitochondrial sur d'autres plateformes

Figure 40 – Juxtaposition des profils de couverture obtenus pour les runs test publiés par *Life Technologies* et dans notre étude

Juxtaposition des profils de couverture provenant d'une part de l'article de blog présenté, et d'autre part des données de séquençage obtenues dans le cadre de l'étude GENESIS. La courbe noire est la couverture moyenne pour les 436 échantillons de GENESIS. Les courbes rouge et bleue correspondent aux couvertures obtenues pour les deux runs de séquençage commentés dans l'article. Afin de mieux visualiser les courbes de manière distincte, l'échelle des ordonnées est indépendante pour les deux expériences. Les courbes sont cependant correctement alignées sur l'échelle des abscisses.

de séquençage ne donne pas un profil de couverture similaire[298]. À la vue de ces éléments, le profil de couverture observé semble donc être plateforme-séquence dépendant. Ainsi, le profil de couverture des données de séquençage du génome mitochondrial effectué sur une plateforme Ion Torrent sera vraisemblablement toujours similaire aux profils représentés sur la figure 40.

1157 variants ont été identifiés en comparaison de la séquence de référence du génome mitochondrial rCRS. Tous les variants fréquents, c'est à dire ayant été observés avec une fréquence supérieure à 5% dans l'étude, sont des variants déjà connus, et référencés dans la base de données MITOMAP. La majorité de ces variants sont des polymorphismes communs du génome mitochondrial. Cependant, certains de ces variants ont déjà été associés avec une hausse du risque de cancer du sein[225] : rs3937033 ou T16519C est observé dans 309 des 436 échantillons inclus dans l'étude, soit 70%, rs2853826 ou A10398 dans 72 échantillons, soit 16%, et rs200191755 ou G9055A dans 40 échantillons, soit 9%. rs62581341 ou T16362C a été observé dans 13 échantillons ; ce SNP a auparavant été spécifiquement associé avec le risque de cancer du sein familial[299].

Parmi les 35 variants, dont 3 insertions-délétions, qui ne sont pas référencées dans MITOMAP, aucun n'a été observé chez plus de 3 individus parmi les 436 inclus. À l'examen visuel des données de séquençage après alignement, aucun élément ne laisse supposer que ces variants seraient des faux-positifs. Ces variants pourraient tout à fait être des variants rares ou des mutations privées (mutations apparues récemment et peu répandues dans la population générale, observées dans quelques familles uniquement). 4 de ces variants sont prédits comme délétères par SIFT et probablement dommageables par PolyPhen. D'autres part, ces variants ne sont référencés dans aucune publication scientifique.

24 variants sont prédits comme à la fois « délétères » par SIFT et « probablement dommageables » par PolyPhen. Parmi eux, aucun n'a été observé dans plus de 3 échantillons, à l'exception du polymorphisme rs2853506 ou A15218G, qui a été observé chez 10 individus,

soit 2% des échantillons séquencés. Ce polymorphisme a déjà été associé avec l'épileptogénèse (développement progressif des symptômes associés à l'épilepsie)[300], mais n'a jamais été évoqué directement dans le cadre du cancer du sein. Ce polymorphisme est situé sur le gène *MT-CYB* codant pour le cytochrome B. Avec le cytochrome c1 et la protéine Rieske, le cytochrome est l'une des trois sous-unités formant le complexe III de la chaîne respiratoire. Certaines altérations de ce complexe ont été directement liées à la modification de ses capacités catalytiques[301].

Pour chacun des gènes mitochondriaux autres que ceux codant pour des ARNs de transfert, l'enrichissement global en variants distincts r_g et l'enrichissement en variants pondéré r_w ont été calculés. Alors que la moyenne de l'enrichissement global r_g est de 64.2 variants/Mb, deux gènes se distinguent avec des valeurs nettement inférieures aux autres : *MT-RNR1* et *MT-RNR2*, avec des valeurs de 35.6 et 37.2 variants/Mb. Ces deux gènes codent respectivement pour les sous-unités 12S et 16S des ARNs ribosomiques, des composants structuraux de la petite et de la grande sous-unité des mitoribosomes. À l'instar des ribosomes cytoplasmiques, les mitoribosomes sont nécessaires à l'assemblage des polypeptides de la mitochondrie. Étant donné leur rôle structurel essentiel, il n'est pas surprenant de constater que ces gènes sont plus conservés que les autres gènes mitochondriaux. De plus, deux gènes codant pour des protéines se distinguent par un taux d'enrichissement global légèrement supérieur à celui des autres gènes : *MT-ATP6* et *MT-CYB* avec un taux d'enrichissement respectif de 86.5 et 84.1 variants/Mb. Un test de Shapiro ayant permis de vérifier que la distribution des valeurs du taux d'enrichissement global pour l'ensemble des gènes codants suit bien une loi normale, les valeurs de ces deux gènes se situent respectivement au niveau des 93$^{\text{ème}}$ et 91$^{\text{ème}}$ percentiles de la distribution observée. Ces deux gènes sont de même caractérisés par un taux d'enrichissement pondéré élevé, puisqu'ils ont deux des valeurs les plus élevées (0.92 et 0.91 variants.individu/Mb). Le maximum observé pour le taux d'enrichissement pondéré est de 1.07 pour le gène *MT-RNR1*. Or, ce gène est parmi les plus conservés puisqu'il présente un taux d'enrichissement global de 35.6 variants/Mb. Cette valeur s'explique par le fait que le génome de référence utilisé provient d'un individu ayant réellement existé, et ayant lui-même porté des polymorphismes, dont deux polymorphismes rares sur *MT-RNR1*. Ainsi, la majeure partie de la population générale, et environ 98% des individus de notre étude porte l'allèle fréquent de ces SNPs, mais c'est l'allèle rare qui est inclus dans la séquence de référence rCRS. À ces positions, un variant est donc détecté pour 98% des échantillons, ce qui augmente fortement la valeur de r_w pour ce gène.

Alors que la technologie Ion Torrent est connue pour être très fiable pour détecter les substitutions, la fiabilité de détection des insertions/délétions est encore sujet à débat. Il apparaîtrait qu'une importante proportion des indels détectées à partir de données générées par Ion Torrent soient des faux-positifs[302–304]. Dans notre étude, seulement 10 des 111 indels initialement détectées ont passé les filtres imposés avec succès. 93% des indels ne passant pas les filtres ont été rejetées dans au moins un échantillon car elles présentaient une répartition déséquilibrée des lectures attestant de leur présence sur les deux brins d'ADN, avec moins de 10% de ces reads sur l'un des deux brins (biais de brin). Ce type de biais a déjà été mis en évidence sur des données issues d'un séquenceur Ion Torrent de manière spécifique, notamment pour les délétions[303]. En outre, parmi nos résultats, 8 des 10 délétions passant les critères de qualité sont situées sur des régions homopolymériques, ce qui les rend moins vraisemblables. Seules les 2 délétions restantes semblent être réellement fiables à l'issue d'un examen visuel des données avec un logiciel de type IGV[305,306]. La technologie Ion Torrent est en effet connue pour être sujette aux erreurs

de séquençage au niveau des homopolymères de taille supérieure à 3, l'incertitude de mesure de la variation de pH lors de la polymérisation de plusieurs nucléotides consécutifs conduisant à une mésestimation de la taille de l'homopolymère associé. Loman et ses collaborateurs[307] ont ainsi estimé le taux d'erreurs d'indels associé aux homopolymères à 1.5 pour 100 paires de bases. De meilleurs résultats concernant l'appel des variants sont généralement obtenus à partir de données de séquençage Ion Torrent en utilisant un algorithme d'alignement basé sur l'algorithme BWT - *Burrows-Wheeler Transform* -, un algorithme d'alignement sur lequel est basé BWA (l'algorithme d'alignement utilisé dans cette étude) ou encore l'algorithme inclus dans la suite d'analyse *NextGene* (SoftGenetics, State College, PA, USA)[308].

Filtrer les variants non-homozygotes nous permet d'obtenir une liste restreinte de variants probables en ayant éliminé une large proportion de faux-positifs générés par le bruit introduit dans les données par le séquençage Ion Torrent. Cependant, ce faisant on choisit de ne pas tenir compte de la potentielle hétéroplasmie sous-jacente, ni des éventuelles mutations somatiques présentes dans les échantillons sanguins analysés, mais plutôt de se concentrer sur l'analyse des variants majoritaires. Cela étant dit, l'étude de l'hétéroplasmie sur des échantillons sanguins ne semble pas extrêmement pertinente dans le cadre de l'analyse du risque de cancer du sein. Il aurait été plus légitime de vouloir étudier l'existence d'un certain degré d'hétéroplasmie au sein des tissus mammaires.

L'objectif de cette étude est d'identifier de potentiels variants pathologiques au sein du génome mitochondrial qui, du fait de leur localisation, auraient été manqués par les technologies classiques de détection, et ce chez des femmes présentant une forte histoire familiale de cancer du sein ne portant pas de mutation pathogène sur *BRCA1* ou *BRCA2*. Étant donné les faibles performances des études de liaison familiale et des études pangénomiques sur le génome mitochondrial, une approche ciblée était nécessaire. Nous avons identifié 1 157 variants, dont certains ont déjà été associés au risque de cancer du sein auparavant. Cependant, nous n'avons pas identifié de profil génomique pathogène flagrant chez ces 436 femmes inclues dans notre étude. Il est donc peu probable que les mutations observées sur le génome mitochondrial permettent d'expliquer l'excès de risque de cancer du sein que ces femmes présentent.

Identifier et comprendre la part manquante de l'héritabilité du cancer du sein nécessite aujourd'hui de mettre au point de nouvelles approches. Les technologies les plus récentes développées peuvent aider à atteindre cet objectif, et le séquençage à haut débit en est une des plus prometteuses. De nouveaux consortiums émergent et permettent de mutualiser les efforts effectués et les capacités, à l'instar de COMPLEXO, dont l'objectif est d'identifier l'héritabilité manquante du risque de cancer du sein en explorant méticuleusement l'exome humain par des approches de séquençage à haut-débit[289].

II. Étude complémentaire 1 : comparaison des performances de deux algorithmes d'alignement

II.1 Objectifs

Une étude complémentaire au travail présenté ici a été réalisée afin de déterminer quel algorithme était le plus plus adapté pour l'alignement des reads issus du séquençage du génome mitochondrial par Ion Torrent. En effet, l'algorithme BWA est un algorithme robuste, ayant fait ses preuves, et en constante amélioration. Cependant, il a été initialement conçu pour l'alignement de données de séquençage issues de la plateforme Illumina, et non Ion Torrent. Ces deux plateformes ne sont pas basées sur le même principe de séquençage, la première détectant les variations de fluorescence suite à l'incorporation d'un nucléotide, alors que l'autre détecte des variations de pH. Or, des technologies différentes peuvent induire des biais de séquençage différents, et ces biais peuvent influencer l'alignement. Par exemple, une des faiblesses de la technologie d'Ion Torrent est la détection du nombre exact de bases constituant un homopolymère, soit une séquence constituée de répétitions de la même base azotée. Par exemple, la séquence suivante CTG**AAAAA**GA contient un homopolymère formé d'une répétition de A. Lorsque la portion d'ADN séquencée contient une répétition d'une même base, la technologie Ion Torrent ne détectera qu'un seul pic de variation de pH pour l'ensemble de l'homopolymère, alors que la technologie d'Illumina, détectera un signal par nucléotide. Ion Torrent est donc beaucoup moins précis qu'Illumina pour déterminer quel nombre exact de nucléotides contient l'homopolymère. De type de biais peut avoir un impact sur l'alignement, et donc au bout du processus, sur la liste des variants détectés. C'est pourquoi Ion Torrent a mis au point son propre algorithme d'alignement, appelé TMAP, qui est disponible publiquement. L'objectif de cette étude complémentaire est de déterminer lequel des deux algorithmes d'alignement BWA et TMAP est le plus approprié afin d'analyser nos données.

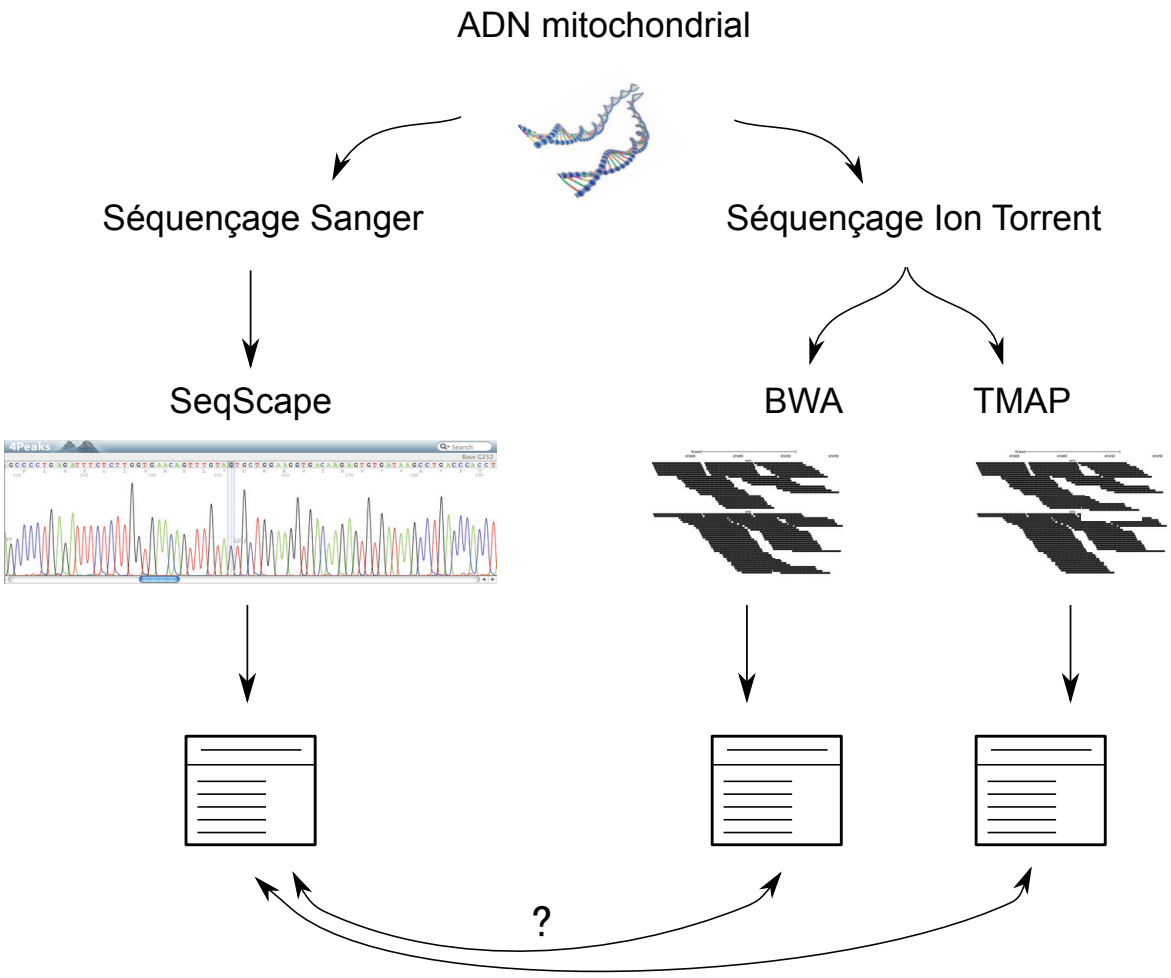
II.2 Principe

Afin d'atteindre cet objectif, j'ai comparé la liste des variants détectés par séquençage Ion Torrent en utilisant l'un ou l'autre des deux algorithmes d'alignement évoqués (BWA ou TMAP) avec les résultats obtenus sur les mêmes échantillons avec la méthode de séquençage la plus fiable à l'heure actuelle, le séquençage Sanger (Figure 41).

Le génome mitochondrial de deux individus a été séquencé par séquençage Sanger au sein de l'équipe. Les analyses de ces données de séquençage Sanger ont été effectuées à l'aide du logiciel SeqScape, et une liste de variants par rapport au génome de référence a été établie pour le génome mitochondrial de ces deux individus. Les pics de fluorescence de chacun des variants détectés par SeqScape ont été vérifiés visuellement. On considère par la suite que ces variants sont réels et représentent la réalité.

Le génome de ces deux individus a d'autre part été séquencé par Ion Torrent, selon le protocole décrit précédemment. Les reads ont ensuite été alignés avec les deux algorithmes d'alignement de manière indépendante. La suite du pipeline d'analyse (réalignement local autour des insertions/délétions, appel des variants, filtrage) décrit en section I.2 .1 a été appliquée de manière strictement identique par la suite.

Figure 41 – Démarche mise en place : comparaison des résultats obtenus avec BWA et TMAP avec ceux issus du séquençage Sanger



II.3 Résultats

Les résultats de l'appel des variants sont présentés dans la Table 23.

Table 23 – Effectif des variants détectés selon l'algorithme d'alignement utilisé

	Total	SNPs	Insertions	Délétions
Individu 1				
Sanger	27	24	2	1
IonTorrent + BWA	25	24	1	0
IonTorrent + TMAP	25	24	1	0
Individu 2				
Sanger	36	33	1	2
IonTorrent + BWA	34	33	1	0
IonTorrent + TMAP	32	32	0	0

Individu 1 Les variants détectés par le pipeline d'analyse utilisant BWA ou TMAP à partir des données de séquençage Ion Torrent sont strictement identiques : 24 SNPs, 1 insertion, 0 délétion. Ces 25 variants détectés par séquençage Ion Torrent figurent bien parmi les 27 variants détectés en Sanger. Ces résultats ne nous permettent pas de discriminer l'un ou l'autre des algorithmes.

Les performances du pipeline d'analyse utilisant BWA ou TMAP peuvent être qualifiées par leur sensibilité et spécificité, en prenant comme référence les résultats du séquençage Sanger. Le logiciel utilisé pour effectuer l'appel des variants, GATK-*UnifiedGenotyper*, construit à chaque position du génome analysé un modèle de vraisemblance. On considère donc comme vraie l'hypothèse selon laquelle l'appel d'un variant est effectué de manière indépendante de l'appel des autres variants.

En réalité, à cause de contraintes techniques, seules 14 943 positions sur les 16 569 positions constituant le génome mitochondrial ont effectivement été séquencées à la fois en Sanger et par Ion Torrent. Sur cette base, on peut alors travailler sur la table de contingence des vrais positifs (VP), faux positifs (FP), vrais négatifs (VN) et faux négatifs (FN) pour chacun des deux algorithme d'alignement utilisés.

Table 24 – Table de contingence des évènements détectés pour BWA et TMAP

Mapper	**VP**	**FP**	**VN**	**FN**
BWA	25	0	14 916	2
TMAP	25	0	14 916	2

La sensibilité d'une méthode correspond à la probabilité de conclure à un vrai positif, c'est à dire à la probabilité de conclure qu'un variant existe sachant qu'il existe réellement. Elle se calcule de la manière suivante :

$$sen = \frac{VP}{VP + FN}$$

La spécificité d'une méthode correspond à la probabilité de conclure à un vrai négatif, c'est à dire la probabilité de conclure qu'un variant n'existe pas sachant qu'il n'existe effectivement pas en réalité. Elle se calcule de la manière suivante :

$$spe = \frac{VN}{VN + FP}$$

On obtient donc pour chacun des deux logiciels d'alignement une sensibilité de 0.962 et une spécificité de 1 (Table 25).

Table 25 – Sensibilité et spécificité de la méthode d'analyse pour l'individu 1

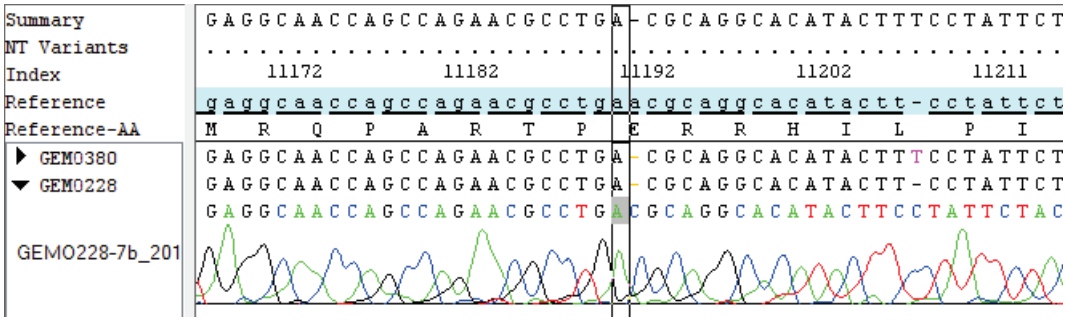
Algorithme	**Sensibilité**	**Spécificité**
BWA	0.962	1.00
TMAP	0.962	1.00

En conclusion, l'analyse des données de ce premier individu ne permet pas de conclure quant à l'algorithme d'alignement le plus approprié.

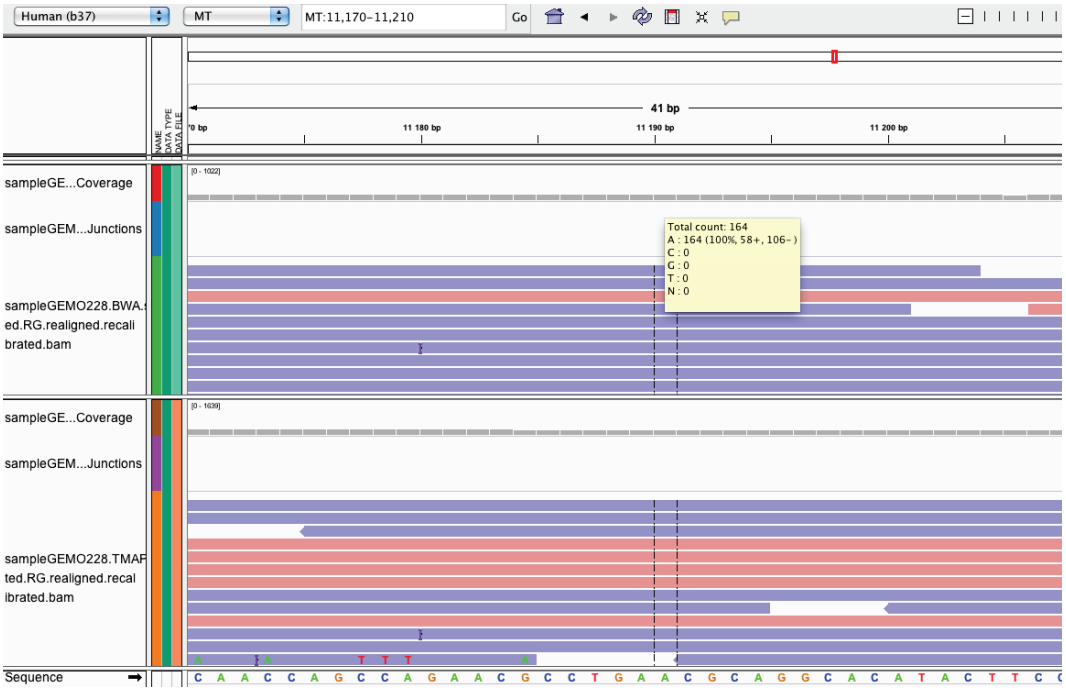
Les deux variants détectés uniquement en Sanger et manqués en Ion Torrent sont respectivement une délétion d'une base A en position 11190 du génome mitochondrial, notée 11190delA, et une insertion d'une base G en position 15049 du génome mitochondrial, notée 15049insG. La visualisation détaillée des pics de fluorescence du séquençage Sanger et du résultat de l'alignement avec BWA et TMAP sont présentées en Figures 42 et 43.

Pour le variant 15049insG, bien que sur la figure 43 un pic noir correspondant à une base G soit visible à cette position, la saturation de la fluorescence correspondant aux bases voisines C pourrait laisser penser que ce variant est un faux positif. Cependant, la visualisation des pics de fluorescence pour le variant 11190delA montre clairement un seul pic vert correspondant à la base A, alors que deux A successifs sont présents dans la séquence de référence. A contrario, la visualisation de l'alignement des reads par BWA ou TMAP ne montre aucune divergence des reads par rapport au génome de référence. Aucun élément n'explique cette différence dans les données au niveau de la position 11 190.

FIGURE 42 – Visualisation des pics de fluorescence obtenus par Sanger et de l'alignement des reads obtenus par séquençage Ion Torrent avec BWA et TMAP pour l'individu 1 à la position 11190



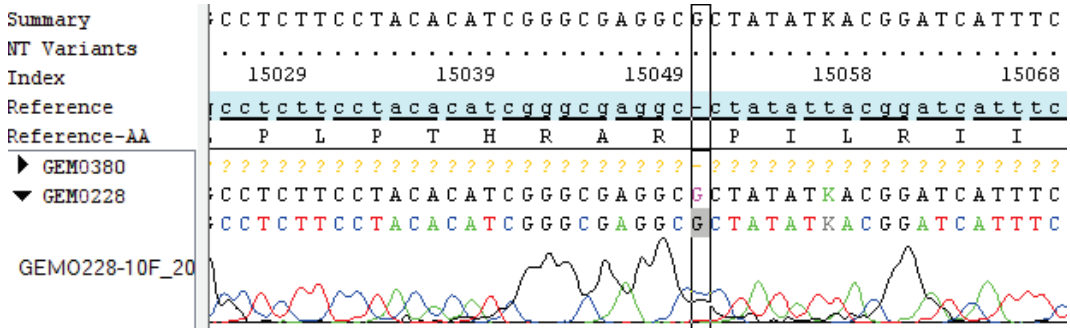
(a) Visualisation des pics de fluorescence du séquençage Sanger



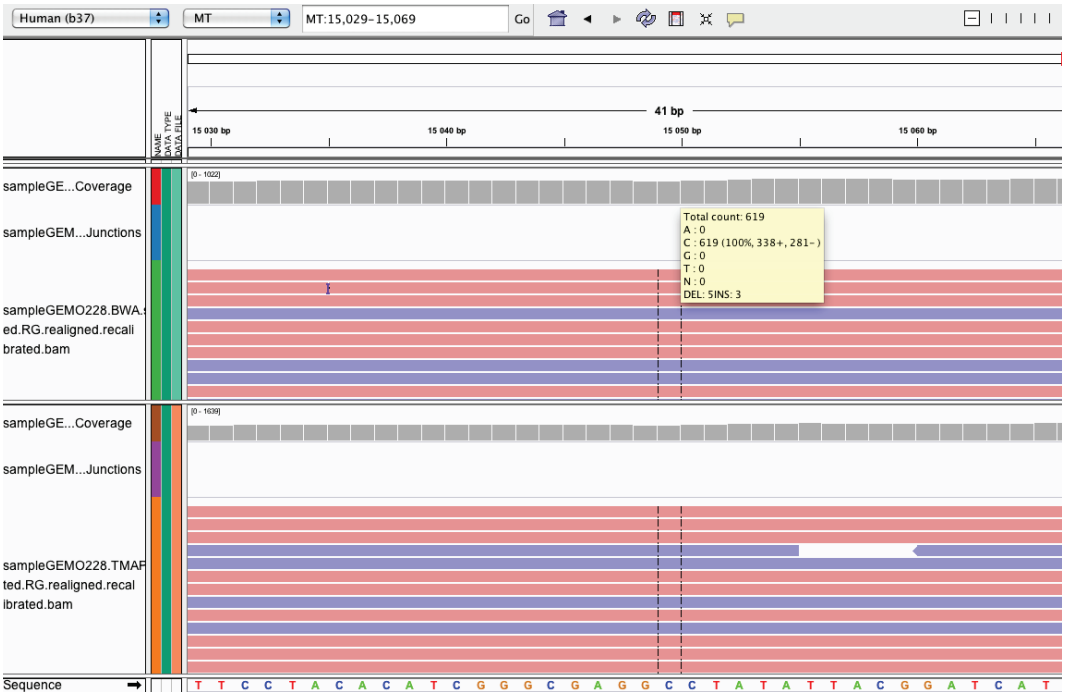
(b) Alignement des reads avec BWA et TMAP

(a) La séquence du génome de référence est représentée sur fond bleu. La position 11190 est encadrée en noir. La séquence obtenue par séquençage est celle représentée tout en haut de la figure (*Summary*). Les pics de fluorescence et les bases correspondantes sont représentés avec leur couleur respective. Alors que deux A sont présents dans la séquence de référence, on observe effectivement un seul pic vert correspondant à un seul A pour cet échantillon.
(b) Visualisation sous IGV (*Integrative Genome Viewer)* du résultat de l'alignement des reads de séquençage par BWA (panel du haut) et TMAP (panel du bas). Les bandes horizontales rouges et bleues représentent des reads alignés respectivement selon l'orientation sens et anti-sens sur le génome de référence. Seules les bases divergentes par rapport au génome de référence sont explicitement indiquées. À cette position, les reads s'alignent parfaitement sur le génome de référence, quelque soit l'algorithme d'alignement utilisé.

FIGURE 43 – Visualisation des pics de fluorescence obtenus par Sanger et de l'alignement des reads obtenus par séquençage Ion Torrent avec BWA et TMAP pour l'individu 1 à la position 15 049



(a) Visualisation des pics de fluorescence du séquençage Sanger



(b) Alignement des reads avec BWA et TMAP

(a) La séquence du génome de référence est représentée sur fond bleu. La position 11 190 est encadrée en noir. La séquence obtenue par séquençage est celle représentée tout en haut de la figure (*Summary*). Les pics de fluorescence et les bases correspondantes sont représentés avec leur couleur respective. Bien que la fluorescence des C soit saturée à cette position, le pic G (en noir) se distingue clairement. Cependant, il est possible que ce variant soit le résultat d'un artefact de fluorescence.
(b) Visualisation sous IGV (*Integrative Genome Viewer)* du résultat de l'alignement des reads de séquençage par BWA (panel du haut) et TMAP (panel du bas). Les bandes horizontales rouges et bleues représentent des reads alignés respectivement selon l'orientation sens et anti-sens sur le génome de référence. Seules les bases divergentes par rapport au génome de référence sont explicitement indiquées. À cette position, les reads s'alignent parfaitement sur le génome de référence, quelque soit l'algorithme d'alignement utilisé.

Individu 2 Sur les 36 variants détectés en Sanger, 34 de ces variants ont été détectés dans les données Ion Torrent alignées avec BWA, alors que 32 l'ont été avec alignement par TMAP. Ces deux algorithmes ont respectivement détecté 33 et 32 SNPs sur les 33 détectés par Sanger. Contrairement au pipeline utilisant TMAP, celui utilisant BWA a réussi à identifier l'insertion détectée par Sanger. De la même manière que pour l'individu 1, la table de contingence des Vrais Positifs (VP), Faux Positifs (FP), Vrais Négatifs (VN) et Faux Négatifs (FN) (Table 26) nous permet de calculer la sensibilité et la spécificité de la méthode d'analyse appliquée avec chacun des deux algorithmes d'alignement (Table 27). Pour cet échantillon, seules 15 081 positions ont effectivement été séquencées à la fois par Sanger et Ion Torrent. On obtient pour les deux méthodes d'alignement une spécificité identique égale à 1 ; pour TMAP une sensibilité égale à 0.889 et pour BWA une valeur de 0.944.

Table 26 – Table de contingence des évènements détectés pour BWA et TMAP

Mapper	**VP**	**FP**	**VN**	**FN**
BWA	34	0	15045	2
TMAP	32	0	15045	4

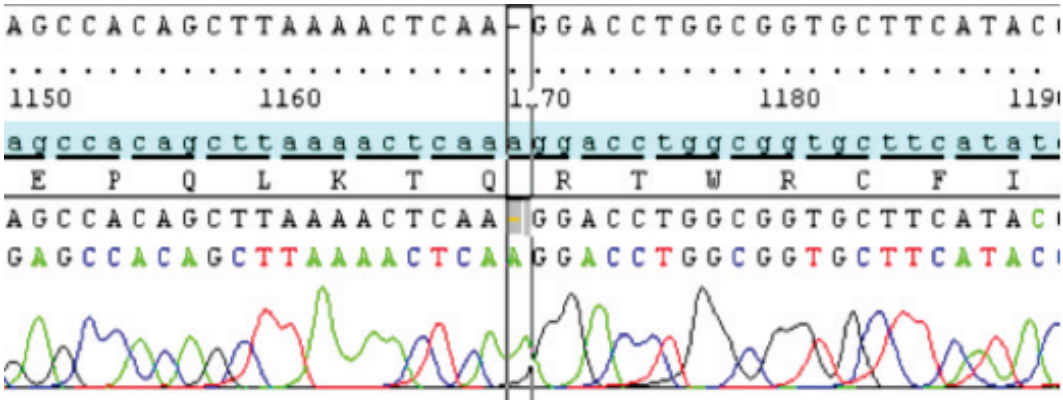
Table 27 – Sensibilité et spécificité de la méthode d'analyse pour l'individu 2

Algorithme	**Sensibilité**	**Spécificité**
BWA	0.944	1.00
TMAP	0.889	1.00

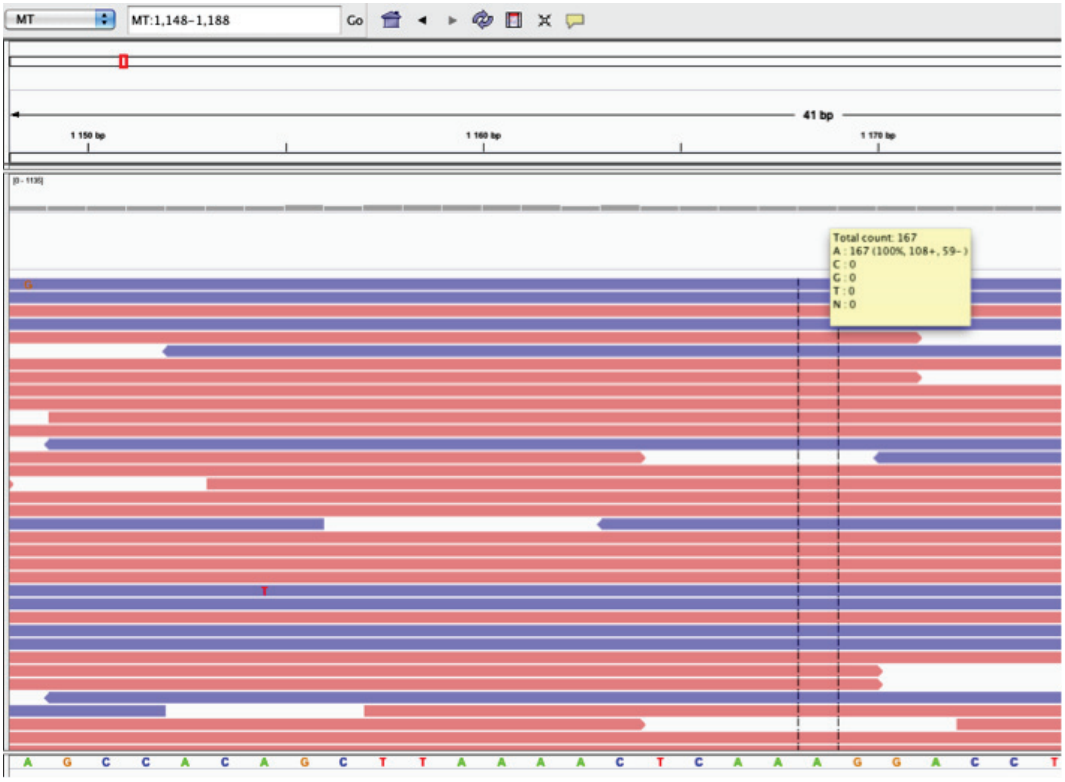
Ainsi, d'après la comparaison de la performance des algorithmes d'alignement sur ce second échantillon, BWA a une meilleure sensibilité que TMAP. De plus, BWA a été capable de détecter correctement une insertion, contrairement à TMAP.

2 variants observés en Sanger n'ont pas été détectés par le pipeline utilisant BWA : une délétion d'une base a en position 1 165, notée 1165delA, et la même délétion que pour l'individu 1 notée, 11190delA. Les visualisations détaillées des profils de fluorescence et d'alignement avec BWA sont présentés dans les figures 44 et 45. Pour ces deux délétions, les profils de fluorescence sont très nets. Cependant, rien dans les données de séquençage Ion Torrent ne laisse supposer l'existence de ces deux délétions.

FIGURE 44 – Visualisation des pics de fluorescence obtenus par Sanger et de l'alignement des reads obtenus par séquençage Ion Torrent avec BWA pour l'individu 2 à la position 1 165



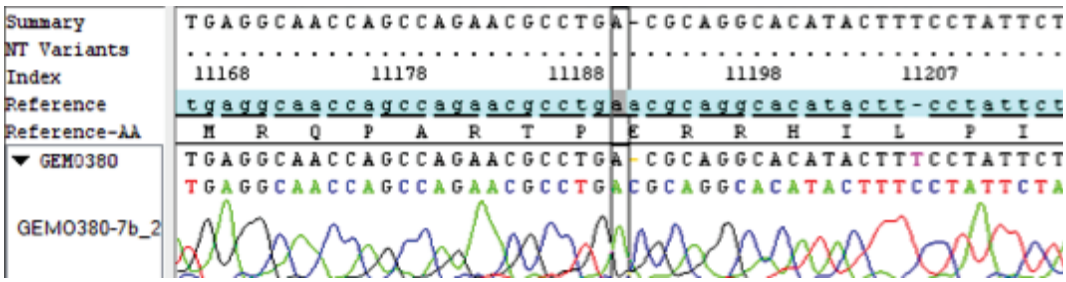
(a) Visualisation des pics de fluorescence du séquençage Sanger



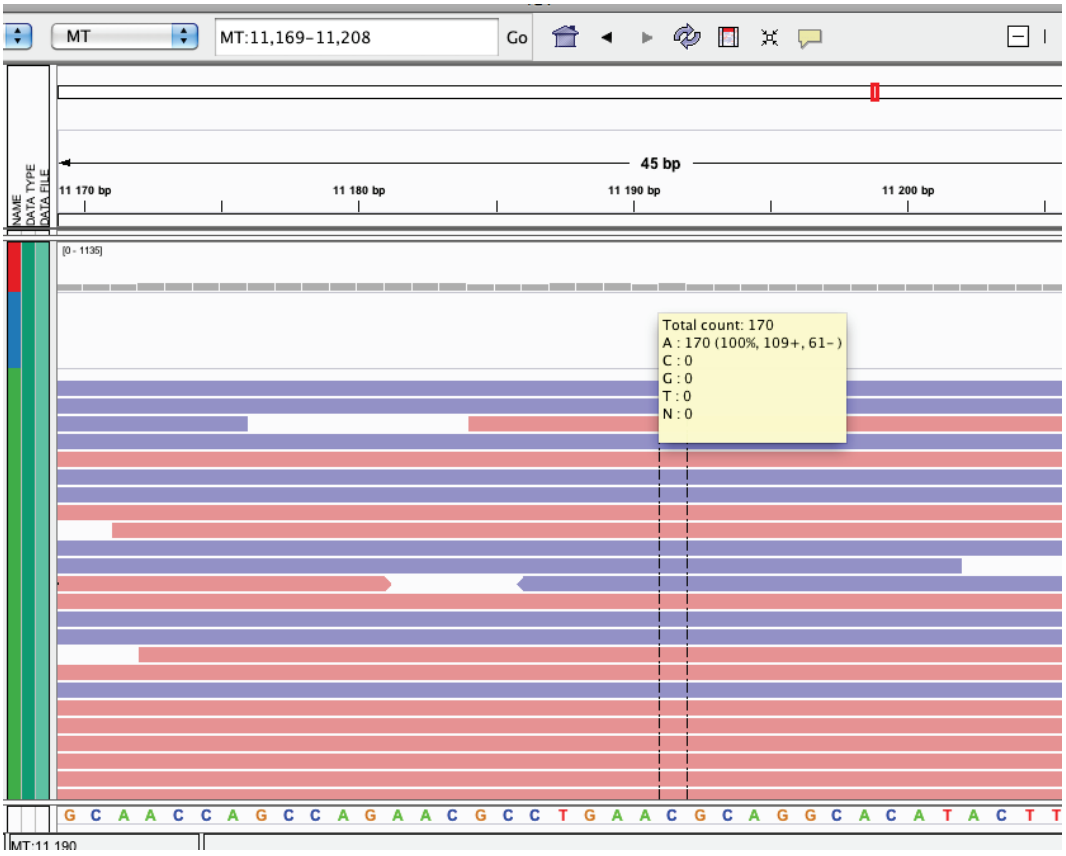
(b) Alignement des reads avec BWA

(a) La séquence du génome de référence est représentée sur fond bleu. La position 1 165 est encadrée en noir. La séquence obtenue par séquençage est celle représentée tout en haut de la figure. Les pics de fluorescence et les bases correspondantes sont représentés avec leur couleur respective. Alors que trois A sont présents dans la séquence de référence, on observe effectivement deux pics verts correspondant à deux A pour cet échantillon.
(b) Visualisation sous IGV (*Integrative Genome Viewer)* du résultat de l'alignement des reads de séquençage par BWA. Les bandes horizontales rouges et bleues représentent des reads alignés respectivement selon l'orientation sens et anti-sens sur le génome de référence. Seules les bases divergentes par rapport au génome de référence sont explicitement indiquées. À cette position, les reads s'alignent parfaitement sur le génome de référence, et aucune délétion n'est détectée.

FIGURE 45 – Visualisation des pics de fluorescence obtenus par Sanger et de l'alignement des reads obtenus par séquençage Ion Torrent avec BWA pour l'individu 2 à la position 11 190



(a) Visualisation des pics de fluorescence du séquençage Sanger



(b) Alignement des reads avec BWA

(a) La séquence du génome de référence est représentée sur fond bleu. La position 11 190 est encadrée en noir. La séquence obtenue par séquençage est celle représentée tout en haut de la figure (*Summary*). Les pics de fluorescence et les bases correspondantes sont représentés avec leur couleur respective. Alors que deux A sont présents dans la séquence de référence, on observe effectivement un seul pic vert correspondant à un seul A pour cet échantillon.
(b) Visualisation sous IGV (*Integrative Genome Viewer)* du résultat de l'alignement des reads de séquençage par BWA. Les bandes horizontales rouges et bleues représentent des reads alignés respectivement selon l'orientation sens et anti-sens sur le génome de référence. Seules les bases divergentes par rapport au génome de référence sont explicitement indiquées. À cette position, les reads s'alignent parfaitement sur le génome de référence, quelque soit l'algorithme d'alignement utilisé.

En conclusion, la méthode de séquençage Sanger reste la méthode de référence pour son exactitude en séquençage, bien que son rendement soit beaucoup plus faible, sans commune mesure avec celui d'Ion Torrent. Les données de séquençage Ion Torrent, quel que soit le pipeline d'analyse qui leur est appliqué, ne permettent pas encore d'atteindre le même niveau de précision pour l'appel des variants. L'analyse des données de séquençage du génome mitochondrial de l'individu 1 a fourni des résultats qui ne nous permettent pas de conclure. Cependant, d'après les résultats de l'analyse de ces même données pour l'individu 2, BWA semblerait plus sensible que TMAP (sensibilités respectivement égales à 0.94 et 0.89) pour détecter les variants du génome mitochondrial. De plus, la spécificité des deux méthodes testées a été estimée à 1 : aucun faux-positif n'a été détecté. Cette étude complémentaire justifie donc le choix de BWA comme algorithme d'alignement pour l'étude des données de séquençage mitochondrial effectuée au sein de la cohorte GENESIS. Afin de confirmer la tendance détectée ici, une étude plus importante serait nécessaire, afin de comparer les performances des deux algorithmes testés sur de larges banques de séquences.

III. Étude complémentaire 2 : Extraction des blocs hautement conservés du génome mitochondrial

III.1 Objectifs

Une seconde étude complémentaire a été réalisée afin de déterminer quelles positions du génome mitochondrial sont strictement conservées au cours de l'Évolution entre plusieurs espèces d'eucaryotes supérieurs. En effet, si certaines positions du génome mitochondrial sont conservées de manière forte, alors les allèles observés à ces positions ont de fortes chances d'être soumis à une pression de sélection purifiante, comme cela a déjà été prouvé pour le génome mitochondrial[309] Cette pression de sélection inhiberait la propagation d'allèles alternatifs à ces positions au sein de la population générale ; les changements apparaissant dans la séquence conservée seraient donc contre-sélectionnés. Si c'est effectivement le cas, alors des mutations affectant ces allèles conservés auront probablement des conséquences fonctionnelles pouvant altérer les fonctions mitochondriales ou leur régulation.

III.2 Principe

La séquence du génome mitochondrial de 8 espèces en plus de l'espèce humaine a été extraite depuis Genbank. La Table 28 répertorie les espèces sélectionnées.

Table 28 – Dénomination et numéro d'accession des espèces sélectionnées pour l'analyse

Espèce	Nom usuel	Identifiant Genbank	Taille (bp)
Ursus americanus	Ours américain	AF303109.1	16841
Mus Musculus	Souris	NC_005089.1	16299
Gallus gallus gallus	Poulet	NC_007236.1	16785
Bos taurus	Vache	GU947010.1	16340
Bufo gargarizans	Crapaud	NC_008410.1	17277
Cynoglossus semilaevis	Sole-langue	NC_012825.1	16731
Lemur catta	Lémurien	AJ421451.1	17036
Oncorhynchus keta	Saumon	NC_017838.1	16656
Homo sapiens	Homme	NC_012920.1	16569

Un alignement multiple de l'ensemble des séquences du génome mitochondrial de ces 9 espèces a été réalisé à l'aide du logiciel Clustal Omega[310,311] v1.2.0. À partir de cet alignement multiple, on effectue la recherche de Blocs Hautement Conservés, notés BHC. On utilise pour cela un algorithme appelé GBlocks, v0.91b[312]. Cet algorithme définit des blocs de séquences conservées à partir de critères simples, tels que le pourcentage de résidus identiques entre toutes les séquences à une position donnée de l'alignement multiple, la proximité avec des positions non-conservées entre espèces, la longueur minimale d'un bloc, la présence de gaps. L'algorithme est très précisément décrit dans la publication associée[312]. Plusieurs paramètres sont modifiables

afin d'ajuster la stringence de définition des BHCs. L'algorithme implémenté dans GBlocks consiste en 6 étapes (la valeur par défaut des seuils est indiquée par l'indice $_{def}$) :

1) Détermination du degré de conservation d_c à chaque position de l'alignement multiple. Ainsi, si on appelle N le nombre total de séquences composant l'alignement multiple, alors chaque position peut être caractérisée de la manière suivante :
   - *non-conservée* si le nombre de bases identiques n_i entre les N séquences alignées est tel que $n_i < IS$, avec $IS_{def} = 0.5 * N + 1$, ou si un gap est observé à cette position.
   - *conservée* si $IS \leq n_i < FS$, avec $FS_{def} = 0.85 * N$,
   - *hautement conservée* si $n_i \geq FS$.
2) Les ensembles de positions consécutives non-conservées dont la longueur l est telle que $l > CP$ sont rejetées. $CP_{def} = 8$.
3) Parmi les blocs restants, on élimine les positions flanquantes jusqu'à ce que chaque bloc soit encadré par des positions hautement conservées. Cela permet d'identifier chaque bloc de manière robuste, avec des bornes stables, et facilite ensuite les éventuels alignements.
4) À cette étape, un bloc doit avoir une longueur minimale L telle que $L \geq BL1$, avec $BL1_{def} = 15$. On évite ainsi de sélectionner des blocs trop courts, pour lesquels la conservation peut être moins robuste.
5) Les positions présentant des gaps sont alors supprimées, ainsi que leurs positions adjacentes si elles sont non-conservées.
6) Enfin, après avoir rejeté les positions présentant des gaps, les blocs restants doivent avoir une longueur minimale $L \geq BL2$, avec $BL2_{def} = 12$.

Cet algorithme a été appliqué sur l'alignement multiple réalisé à partir des séquences d'ADN mitochondrial des 9 espèces sélectionnées avec les valeurs de paramètres suivantes :

- $IS = 5$,
- $FS = 8$,
- $CP = 12$,
- $BL2 = 10$.

D'après les résultats obtenus avec cet algorithme, 13 807 positions sur les 16 569 bases qui constituent le génome mitochondrial humain sont considérées comme appartenant à des blocs hautement conservés du génome mitochondrial, soit environ 83%. Cette proportion est très élevée, et l'information que l'on peut en retirer n'est pas très discriminante, puisque quasiment l'intégralité du génome mitochondrial se trouve labellisée *hautement conservée*.

Un autre critère de sélection a donc été utilisé afin de déterminer les positions hautement conservées du génome mitochondrial. Ces positions ont été définies comme celles ayant 100% d'identité entre les 9 séquences comparées, c'est à dire les positions consensus de l'alignement. Bien que cette méthode soit plutôt stringente, il s'avère que 5 488 sur 16 569 positions du génome mitochondrial sont strictement conservées entre les 9 espèces sélectionnées, soit 33%. Ce génome comportant peu de séquences intergéniques et aucun intron, ces positions sont uniformément réparties sur l'intégralité du génome mitochondrial.

Cette seconde étude complémentaire nous a ainsi permis d'établir la liste des 5 488 positions du génome mitochondrial parfaitement conservées entre les 9 espèces comparées. Ces positions peuvent alors être annotées en sortie de l'appel des variants. Un variant génomique affectant une position annotée hautement conservée a donc plus de chances d'avoir des conséquences fonctionnelles importantes.

DISCUSSION

Le séquençage complet du génome humain a constitué une avancée majeure en génomique humaine. Les espoirs et les attentes suscités par cette avancée ont été à la hauteur de l'investissement fourni, à la fois sur le plan financier et humain. Alors que le séquençage du génome humain se démocratise de plus en plus, il a été appliqué à de très nombreuses pathologies dans le but d'élucider leurs origines génétiques. Pour une partie des pathologies étudiées, ce fut un succès. De même, l'approfondissement de nos connaissances du génome humain et de ses variations ont permis au génotypage de devenir de plus en plus précis et exhaustif, permettant dès les années 2 000 la mise en place d'études pangénomiques. Ces nouvelles techniques d'investigation du génome ont, dans une certaine mesure, permis d'identifier d'une part les variants responsables d'un certain nombre de maladies monogéniques, et d'autre part ceux qui augmentent le risque d'apparition d'une pathologie donnée. Mais ces découvertes sont globalement restreintes aux pathologies à l'étiologie « simple ». Seuls les variants fréquents et ayant un effet relativement marqué ont pu être identifiés avec certitude.

Il existe encore de nombreux paramètres liés à la prédisposition génétique à certaines maladies qui restent encore inconnus à l'heure actuelle. Ces pathologies sont dites complexes, ou encore multifactorielles. En effet, leur survenue peut être influencée par de nombreux gènes et variants génomiques distincts, interagissant potentiellement entre eux. De même, l'expression de certains variants génomiques peut être soumise à l'influence de l'environnement ou du style de vie. Enfin, certains de ces variants sont peu fréquents dans la population générale, et il est certain que de nombreux variants rares restent à être identifiés.

Le cancer du sein est un exemple typique de pathologie multifactorielle. Bien qu'une large proportion des cas diagnostiqués soient sporadiques, 5% à 7% des cas de cancer du sein sont familiaux. Le taux d'incidence élevé du cancer du sein dans la population générale (une femme sur 9 sera confrontée à un cancer du sein au cours de sa vie) rend cette proportion non négligeable. De plus, la génétique joue également un rôle dans le risque d'apparition de cancer sporadique, puisque certains variants, sans pour autant avoir une pénétrance de 80% comme certaines mutations sur *BRCA1*, augmentent très légèrement le risque de cancer du sein, de l'ordre de 1% à 3%. C'est pourquoi l'étude de la génomique du cancer du sein est un des sujets de recherche les plus actifs sur ce type de cancer.

Cependant à l'heure actuelle, comme pour la plupart des pathologies multifactorielles, on ne sait expliquer qu'environ 50% de la composante génétique du risque de cancer du sein. La proportion restante de l'héritabilité de cette maladie reste inexpliquée. C'est dans ce contexte que j'ai effectué mes travaux de thèse, en cherchant à déterminer si, dans une certaine mesure, une partie de cette héritabilité manquante peut être expliquée par des variants génétiques localisés non pas sur le génome nucléaire, qui a déjà été longuement étudié, mais sur le génome

mitochondrial, présent en plusieurs dizaines de copies dans quasiment toutes nos cellules, mais parfois un peu laissé pour compte dans le cadre des études génomiques. Dans cet objectif, je me suis intéressée à plusieurs aspects relatifs à cette problématique.

Le premier m'a conduit à étudier les interactions potentielles entre certains variants du génome mitochondrial et du génome nucléaire, ainsi qu'à celles entre un variant mitochondrial donné et le facteur non-génétique lié au style de vie qu'est la consommation d'alcool. Cette étude a été réalisée dans le cadre de la population générale, en prenant en compte aussi bien les cas de cancer du sein sporadiques que familiaux. J'ai modélisé le risque de cancer du sein par régression logistique, une approche classique en épidémiologie génétique. Cette étude avait pour objectif de répliquer des résultats précédemment obtenus, à savoir une première interaction entre 2 polymorphismes, l'un sur le génome mitochondrial, l'autre sur le génome nucléaire, et une seconde interaction entre un polymorphisme mitochondrial et la consommation d'alcool. Les interactions précédemment mises en évidence n'ont pas pu être observées dans cette étude de réplication. Une association potentielle entre un des SNPs étudiés et le risque de cancer de la prostate, également modélisé, a été mise en évidence. Le résultat de cette tentative de réplication illustre bien la nécessité de valider les conclusions scientifiques, et de les étayer de preuves solides et robustes. Cela est parfois difficile dans le cadre de l'étude des liens entre le cancer et l'environnement, le style de vie. Il est parfois difficile de mesurer l'exposition environnementale de manière précise, standardisée, et non-biaisée. Personne ne peut répondre précisément à la question « Combien de grammes d'alcool avez-vous consommé par jour durant les 10 dernières années ? », et le cas échéant, est-il fiable de se baser sur une réponse estimée d'après des souvenirs étalés sur 10 ans ? Il est en outre complexe d'avoir à disposition un second jeu de données indépendant, pour lequel l'exposition environnementale à été mesurée selon des critères similaires à l'étude initiale afin de répliquer les résultats obtenus.

L'étude des interactions entre le risque de cancer et les facteurs environnementaux subis et les facteurs de risques comportementaux est une des priorités de santé publique depuis 2009 en France, dans le cadre du Plan Cancer et du Plan Santé-Environnement. Collecter les informations nécessaires à l'étude des interactions potentielles requiert des moyens techniques, financiers et humains conséquents, notamment en mettant en place des études prospectives dont le principe est d'assurer le suivi d'un échantillon de la population générale de taille élevée. Cela permet de connaître l'exposition des individus inclus dans l'étude avant même qu'ils ne déclarent une maladie. Elles ne sont cependant pas adaptées à toutes les pathologies, en particulier celles pour lesquelles l'exposition responsable de leur apparition a eu lieu de nombreuses années avant que les premiers symptômes n'apparaissent. À l'échelle internationale, le Centre International de Recherche contre le Cancer effectue une veille de la littérature scientifique publiée, et édite régulièrement une liste des carcinogènes potentiels, en les classant en fonction de leur statut, à savoir : agent cancérigène avéré, agent probablement cancérigène, agent potentiellement cancérigène, agent inclassable quant à sa cancérigénicité, ou enfin agent probablement pas cancérigène.

La détection d'interactions impliquant à la fois le génome et l'environnement représente un challenge à la fois sur le plan des études mises en place, nécessitant de très larges échantillons, mais également du point de vue méthodologique. Une étude récente[313] a effectué des simulations afin de déterminer dans quelles conditions les interactions gène-environnement peuvent

être détectées, en particulier dans le cadre d'études GWAS. Ils ont ainsi modélisé un ensemble de traits phénotypiques, contrôlés par un grand nombre de polymorphismes (modèle polygénique), certains ayant un effet plus fort que d'autres sur le trait étudié, ce qu'ils nomment l'architecture génétique. Ils ont ensuite conceptualisé un « changement d'environnement » (qui représenterait, par exemple, une modification des habitudes alimentaires, une vie plus sédentaire, etc...) qui modifie l'architecture génétique, c'est à dire l'effet d'une fraction des polymorphismes ayant un impact sur le trait étudié. Plusieurs modèles ont été estimés, en faisant varier notamment la proportion de la population soumise au nouveau système environnemental par rapport à la fraction qui continue d'être exposé à l'ancien système. Leurs conclusions sont les suivantes : la plupart des interactions n'ont pas été détectées lorsque l'on considère chaque SNP individuellement. À l'inverse, la puissance statistique est suffisante pour détecter les interactions gène-environnement lorsque l'on considère l'ensemble des allèles causaux sous la forme d'une variable qui somme le nombre d'allèles à risque portés, variable appelée score de risque génétique - ou GRS, pour *Genetic Risk Score*. Ils concluent enfin qu'il est très probable qu'on ne puisse pas détecter certaines de ces interactions bien qu'elles influent sur l'architecture génétique sous-jacente. Ainsi, les approches classiques doivent être adaptées au contexte spécifique des interactions gène-environnement, notamment dans le cadre des études GWAS.

Dans l'état de nos connaissances actuelles, la part génétique de la prédisposition est radicalement différente entre le cancer du sein de type sporadique et celui de type familial. En effet, bien que la pénétrance de certaines mutations soit très élevée, par exemple de l'ordre de 80% sur *BRCA1*, il existe encore une proportion non-négligeable de l'héritabilité du cancer du sein familial encore inexpliquée, de l'ordre de 50% à 60%[314]. C'est sur l'étude de la susceptibilité génétique de ces individus diagnostiqués pour un cancer du sein et présentant des antécédents familiaux que la suite de mon travail a porté.

Il existe deux catégories d'individus atteints d'un cancer du sein et ayant des antécédents familiaux : ceux pour lesquels la mutation causale est identifiée, et ceux qui ne portent aucune des mutations causales connues. Je me suis tout d'abord intéressée à la première catégorie car, bien que la mutation à l'origine de leur maladie soit identifiée, ces individus présentent une très forte hétérogénéité en termes de symptômes. Pour certains les symptômes se déclarent à un âge précoce, alors que d'autres à un âge plus avancé ; pour certains, le cancer n'atteint qu'un seul sein, alors pour d'autres il est bilatéral ; certains se voient également diagnostiquer un cancer de l'ovaire, alors que d'autres non ; et dans certaines familles, bien que présentant une mutation causale à forte pénétrance, les individus sont moins affectés par la pathologie que la moyenne. Afin d'étudier si d'autres facteurs génétiques modifient l'association entre les mutations causales localisées sur les gènes *BRCA1/2* et le risque de cancer du sein, j'ai entrepris d'étudier spécifiquement les variations du génome mitochondrial, et en particulier les haplogroupes, en tant que modificateur potentiel. À l'aide d'une approche innovante couplant algorithmique, phylogénie et épidémiologie classique, j'ai identifié, parmi un échantillon conséquent de porteurs avérés de mutations sur le gène *BRCA2*, une branche de l'haplogroupe T, un haplogroupe relativement fréquent dans la population caucasienne, qui est moins enrichie en cas que le reste de la branche. L'haplogroupe T1a1 semblerait donc être un modificateur de l'association entre les mutations causales détectées sur *BRCA2* et le risque de cancer du sein. L'ensemble des éléments issus de la littérature présentés dans le chapitre 2 justifient de formuler l'hypothèse selon laquelle les cellules appartenant à l'haplogroupe T1a1 pourraient avoir de meilleures capacités de résistance

au stress oxydatif, ou y être moins exposé que les cellules appartenant à d'autres haplogroupes. Cette potentielle résistance aux effets du stress oxydatif aurait pour conséquence d'occasionner moins de cassures sur l'ADN, et en particulier de cassures simple-brin, normalement prises en charge en partie par le système de recombinaison homologue. Or, dans le contexte où la protéine BRCA2 n'est pas fonctionnelle, ce système de réparation peut être délaissé au profit d'un système de recombinaison homologue alternatif, appelé aHR. L'usage de ce processus de réparation est d'autre part stimulé par l'inhibition de BRCA2, mais requiert la présence de la protéine BRCA1 fonctionnelle. Cette hypothèse serait cohérente avec le fait que la modification d'effet observée ne l'a été que chez les porteurs de mutations sur *BRCA2*, et non chez les porteurs de mutations sur *BRCA1*. Il a enfin été montré que le processus de réparation aHR favorise la perte d'hétérozygotie et contribue à l'instabilité génomique. Ainsi les cellules T1a1, moins sujettes aux cassures simple-brin, auraient donc moins recours à l'usage du système de réparation aHR, et de ce fait seraient moins susceptibles de développer une instabilité génomique. Les *cybrids*, cellules hybrides permettant de tester spécifiquement les performances cellulaires en fonction de variations du génome mitochondrial, sont devenus un outil fréquemment utilisé afin d'étudier les conséquences fonctionnelles des variants mitochondriaux. La mitochondrie est étudiée dans toutes une variété de pathologies : cardiaques, neurodégénératives, respiratoires, musculaires, ou oncologiques. Le nombre de publications utilisant des *cybrids* ne cesse de croître. Il est indéniable que la confirmation des hypothèses fonctionnelles formulées en amont se fera à travers la mise en place de telles expériences impliquant les *cybrids*.

Enfin, après m'être intéressée aux individus porteurs d'une mutation causale identifiée, j'ai étudié le génome mitochondrial de femmes diagnostiquées pour un cancer du sein, mais ne portant aucune mutation pathogène connue sur les deux principaux gènes de prédisposition au cancer du sein *BRCA1* et *BRCA2*, et ce dans l'objectif de caractériser de potentiels variants de susceptibilité localisés sur le génome mitochondrial. Le génome mitochondrial de 436 femmes non-mutées sur *BRCA1/2* ayant un cancer du sein a été séquencé. Les résultats du séquençage ont été comparés au génome de référence, et plus de 1 150 variants ont été détectés, tous échantillons confondus. Certains variants détectés avec fiabilité, avec une fréquence faible, n'ont jamais été décrits ailleurs, et ne sont pas présents dans la liste des variants connus issue de MITOMAP. 24 variants, dont 2 non-référencés jusqu'alors, sont prédits par PolyPhen comme « probablement dommageables » et comme « délétères » par SIFT. D'autre part, les deux gènes *MT-ATP6* et *MT-CYB* sont les plus fréquemment mutés, en tenant compte à la fois de leur taille et du nombre d'individus porteurs de mutations sur ceux-ci. Les variants mis en évidence par ce travail ne sont que des variants candidats à être étudiés plus en détails, et en aucun cas les arguments présentés ici n'en font des variants de prédisposition. De même, aucune structure génomique commune n'a été observée chez l'ensemble ou une sous-partie de ces femmes. Certains ont tentés d'établir un score de pathogénicité des variants mitochondriaux, en particulier pour les variants affectant les ARN de transfert mitochondriaux[315]. Cependant, dans le contexte de l'étude de la pathogénicité des variants mitochondriaux, tous s'accordent à dire que la méthode de référence afin d'étudier les conséquences de ces variants est l'utilisation des *cybrids*, ou tout du moins que les résultats de tels essais fonctionnels doivent être pris en compte dans l'établissement d'un score de pathogénicité[315,316]. Bien qu'aucun essai fonctionnel n'ai été réalisé dans notre étude, il est cependant intéressant d'avoir caractérisé le génome mitochondrial de ces femmes dans la perspective d'essayer d'identifier des variants contributeurs à l'héritabilité manquante du cancer du sein. De nombreuses études ont vainement tenté d'identifier un candi-

dat *BRCA3*, sans succès jusqu'à présent. Le nombre croissant d'études ayant échoué dans cette quête soulève des questions quant à son existence même, et il apparaît comme étant de plus en plus probable qu'il n'existe pas de *BRCA3*. La part génétique encore non-expliquée du risque de cancer du sein représenterait l'effet combiné de nombreux polymorphismes relativement fréquents ayant une pénétrance allant de faible à moyenne[317], et de quelques variants rares à forte pénétrance. Lynch et al.[314] vont même jusqu'à évoquer la possibilité que la part inexpliquée de la prédisposition au cancer du sein familial ne soit due qu'à des mutations spécifiques de chaque famille. Cette observation est d'une part issue de leurs propres travaux, dans lesquels une mutation à forte pénétrance a été identifiée au sein d'une famille avec des antécédents de cancer du sein familial, mais n'a été retrouvée chez aucun des cas issus de 22 autres familles de profil similaire (négatives pour les mutations connues sur *BRCA1/2*, *PTEN*, et *p53*). D'autre part, les plus récentes études ayant identifié des mutations de prédisposition au cancer du sein familial ont toutes échoué à les valider dans d'autres familles également[318]. Ces observations sont cohérentes avec le fait que le cancer du sein familial se positionnerait plutôt dans un modèle « Même maladie, mais causes différentes ». La mutation causale observée dans une famille donnée pourrait très bien ne pas être retrouvée dans aucune autre famille étudiée. Ainsi, les approches basées sur études populationnelles, telles que les études gène candidat ou les études pangénomiques, conçues pour identifier les origines de maladies dans le cadre d'un modèle de type « Même maladie, mêmes origines », ne peuvent pas détecter ce type de mutations familles-spécifiques. La méthode d'identification de mutations de prédisposition la plus efficace reste donc le séquençage massif de l'exome ou du génome des cas d'une même famille, afin d'isoler les mutations causales potentielles[319–321].

Nous sommes encore bien loin de connaître l'ensemble des facteurs qui influent sur le risque de cancer du sein. Cette maladie multifactorielle peut être reliée à de nombreux facteurs, à la fois génétiques et environnementaux. Alors que les statistiques en font l'une des premières causes de mortalité féminine dans le monde, cette pathologie est en réalité très hétérogène. Le développement de la prise en charge des patients selon une approche de « médecine personnalisée » requiert de tenir compte à la fois du patrimoine génétique de l'individu et de l'existence d'antécédents familiaux, mais également de l'environnement auquel il est exposé, ainsi que de son style de vie. Acquérir une vision globale de cette maladie dans laquelle tous ces aspects seraient intégrés permettra à terme, de réduire le nombre de cas en limitant les expositions et les comportements à risque, d'abaisser la mortalité en surveillant au plus près les individus à haut risque et en les diagnostiquant le plus tôt possible, et enfin de mieux caractériser la maladie à l'échelle de l'individu lorsque celle-ci se déclare, afin d'administrer les traitements les plus susceptibles de fonctionner. De nombreux acteurs travaillent déjà à cette tâche, que ce soit en recherche fondamentale, en recherche clinique, ou en recherche translationnelle.

Il est certain que l'épidémiologie génétique et la bioinformatique continueront de jouer un rôle majeur afin d'atteindre ces objectifs. Ces deux disciplines ont grandement contribué aux avancées les plus poussées dans l'étude de la génétique des pathologies humaines. La bioinformatique est maintenant intégrée dans de nombreux domaines en sciences de la Vie : en génomique médicale bien sûr, avec l'émergence de la médecine personnalisée, la mise en place des études pangénomiques, les puces d'expression génique... Mais la bioinformatique est également présente dans d'autres disciplines. En phylogénie moléculaire, science qui s'attache à l'étude de l'évolution des génomes ; en métagénomique, discipline ayant pour but de carac-

tériser le contenu génétique d'un échantillon issu d'un milieu complexe, ce qui a permis de découvrir de nombreuses nouvelles souches bactériennes et autres espèces ; ou encore en transcriptomique, en métabolomique, et dans l'ensemble des sciences de production de données à haut débit regroupées sous le terme *omics*. La bioinformatique tient également un rôle majeur en protéomique structurale, donc l'objectif est le criblage à haut-débit de molécules afin de déterminer leur structure moléculaire. Quel que soit le domaine des sciences de la Vie, les possibilités révélées par la bioinformatique sont immenses, et le volume de données générées par ces approches n'avait jamais été observé auparavant. Les espoirs suscités par l'incorporation de la bioinformatique appliquée dans certains milieux de recherche étaient tellement grands que la mise en place des infrastructures nécessaires à la synthèse de ces nouveaux types de données s'est réalisée relativement rapidement.

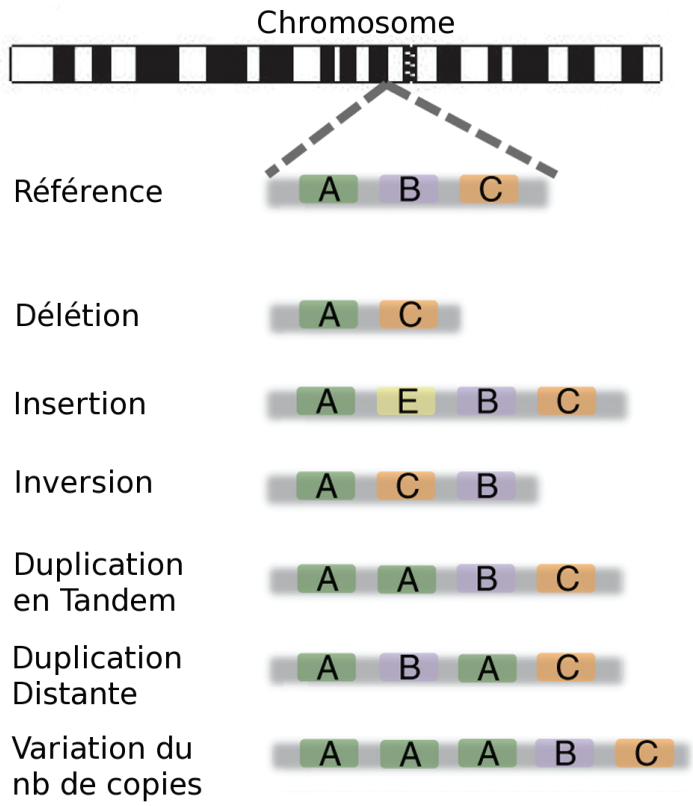
C'est l'analyse des données générées qui représente à l'heure actuelle le goulot d'étranglement au sein du processus global d'analyse. Les analyses les plus simples sont aujourd'hui effectuées en routine : détection de variations génomiques ponctuelles, mesure de l'association de polymorphismes avec un risque donné par régression logistique, mesure de l'expression des gènes, etc... Cependant, chaque secteur fait face à ces propres écueils, et ce pour plusieurs raisons. En premier lieu, le volume de données générées représente un contrainte sous bien des aspects, à la fois pour le stockage, la gestion, le traitement, et la distribution de ces données. En moyenne, un fichier contenant les données de séquençage du génome normal d'un individu à une profondeur de 30X (couverture moyenne) après alignement représente 200 Go, et ce sans tenir compte de l'intégralité des fichiers intermédiaires qui seront nécessairement générés lors de l'analyse de ces données. Déterminer quels sont les variants portés par cet individu par rapport au génome de référence, requiert aujourd'hui en moyenne 360h soit 15 jours complet de temps machine, avec 48 Go de RAM, et 1 unique CPU de calcul. Ce genre d'analyse ne peut être efficacement envisagée sur un simple ordinateur de bureau. Afin de réduire ce temps d'analyse face à ce flux de données, des clusters de calculs ont fait leur apparition, et permettent de paralléliser ces analyses sur un grand nombre de processus. Cependant, le coût d'un tel matériel est parfois rédhibitoire, un cluster de calcul coûtant plusieurs centaines de milliers d'euros. D'autres part, ce cluster de calcul doit pouvoir être utilisé et maintenu, ce qui implique également un investissement en terme de main-d'oeuvre qualifiée. Certaines structures de recherche n'ont clairement pas le budget pour investir dans de telles infrastructures matérielles, et d'autres solutions commencent à émerger. À l'ère des *Big Data*, il est aujourd'hui possible d'effectuer ces analyses sur le *cloud*, réseau virtuel de processeurs connectés gérés par des plateformes telles qu'Amazon ou Microsoft Azure. Le principe est de louer une plateforme de calcul de haute capacité possédant les logiciels d'analyses requis pour un temps donné. Le recours à ce genre de plateforme possède ses propres limitations telles que l'utilisation technique des services désirés en mode de calcul distribué, le respect de certaines limitations lors de son utilisation, la vitesse de transfert des données. Il soulève également des questions éthiques et sur la confidentialité des données. Concernant le transfert de volumes de données extrêmement important via Internet, certaines plateformes telles qu'Amazon proposent de réceptionner les données sur un support physique afin de pouvoir les transférer localement beaucoup plus vite sur leurs grilles de calcul. Le *cloud* reste donc une option destinée aux chercheurs appartenant à des laboratoires de petite taille, ou qui n'ont à analyser ce genre de données qu'épisodiquement. Le développement de la bioinformatique sur le *cloud* est un bon exemple du type d'adaptations exigées à la suite des difficultés rencontrées dans le domaine.

Le second écueil majeur rencontré est l'atteinte de la limite de ce que les méthodologies classiquement utilisées permettent de faire en analyse de données. Je reprend ici deux exemples en relation avec le travail que j'ai réalisé pendant ma thèse, à savoir l'analyse de données de séquençage en génomique du cancer, et l'analyse de données de génotypage issues d'études pangénomiques.

On sait aujourd'hui détecter les variants génomiques d'une seule base - *Single Nucleotide Variant*, SNV - avec une sensibilité et une spécificité proche de 1 dans la plupart des régions du génome humain normal, et ce grâce à des pipelines d'analyses tels que celui que j'ai moi-même mis en place dans la troisième partie de ma thèse. Même dans le cas simple de l'analyse de notre génome constitutionnel, certaines régions restent difficiles à appréhender par séquençage. Par exemple, la présence de régions répétées, contenant des microsatellites ou des motifs homopolymériques, peuvent d'une part provoquer un déraillement de la polymérase, et d'autre part, rendre très complexe la phase d'alignement. En effet, lorsque l'on dispose de reads courts, leur alignement ne peut pas s'effectuer de manière unique sur le génome de référence, et n'est donc pas fiable. L'intérêt des régions microsatellites est pourtant loin d'être négligeable puisqu'ils sont couramment utilisés en sciences forensiques pour l'identification d'individus, et en clinique pour leur implication dans un certain nombre de pathologies, notamment neurologiques. On peut citer à titre d'exemple la maladie de Huntington qui se caractérise par l'expansion d'un triplet CAG[322]. Des méthodes d'analyse spécifiquement dédiées à la détection de ces motifs courts - ou *Short Tandem Repeats*, STR - commencent à voir le jour, notamment certaines basées sur des reads de séquençage plus longs que la moyenne, et utilisant des modèles statistiques plus complexes[323] (Modèles de Markov cachés,etc..). La génomique du cancer fait elle aussi face à des challenges ambitieux. En effet, le génome d'un tissu tumoral est la plupart du temps caractérisé par toute une série d'altérations qui vont de la simple mutation somatique ponctuelle d'une seule base à de complexes réarrangements chromosomiques, souvent accompagnés par des pertes ou des gains de portions chromosomiques. La détection et la caractérisation précise de l'ensemble de ces altérations, et notamment de celles appelées variants structuraux, est un des challenges majeurs aujourd'hui en bioinformatique de la génomique du cancer.

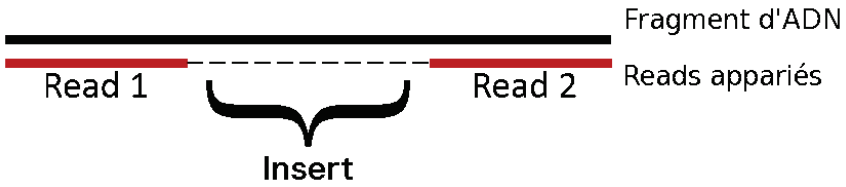
Les variants structuraux regroupent l'ensemble des variations qui altèrent la structure d'une portion chromosomique. Les motifs les plus fréquents de variants structuraux sont schématiquement représentés sur la figure 46.

Figure 46 – Différents types de variants structuraux
adapté d'après Baker et al., 2012 [324]



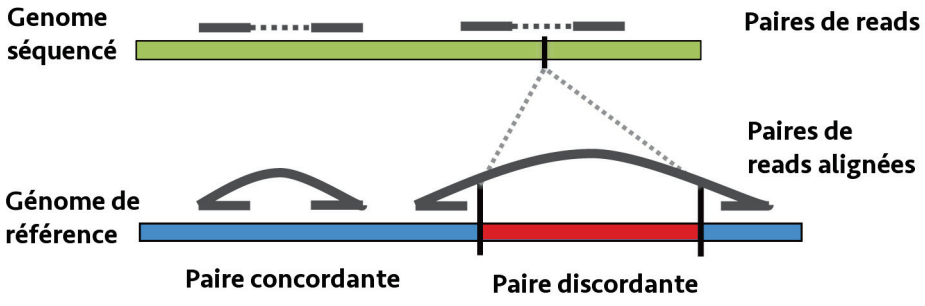
Il existe quatre principales méthodes d'études des variants structuraux. La première consiste à travailler à partir de données issues de séquençage *paired-end*. Dans ce type de séquençage, les reads séquencés correspondent aux deux extrémités d'un même fragment d'ADN de taille donnée, et sont donc appariés deux à deux (Fig. 47). Seules les extrémités du fragment d'ADN sont séquencées, on sait donc en moyenne quelle distance, correspondant à la taille de l'insert non-séquencé, sépare les deux reads d'une même paire.

Figure 47 – Séquençage *paired-end*



En théorie, dans le cas normal, les deux reads d'une même paire doivent donc s'aligner sur le génome de référence en étant éloignés l'un de l'autre de la distance théorique ciblée appelée *insert size*. Or, si une paire de reads est issue d'une portion chromosomique ayant subi un réarrangement structural, la position des deux reads peut ne pas correspondre à leur position attendue. On les appelle alors des paires discordantes. Une tel exemple est représenté en Figure 48 dans le cas d'une délétion.

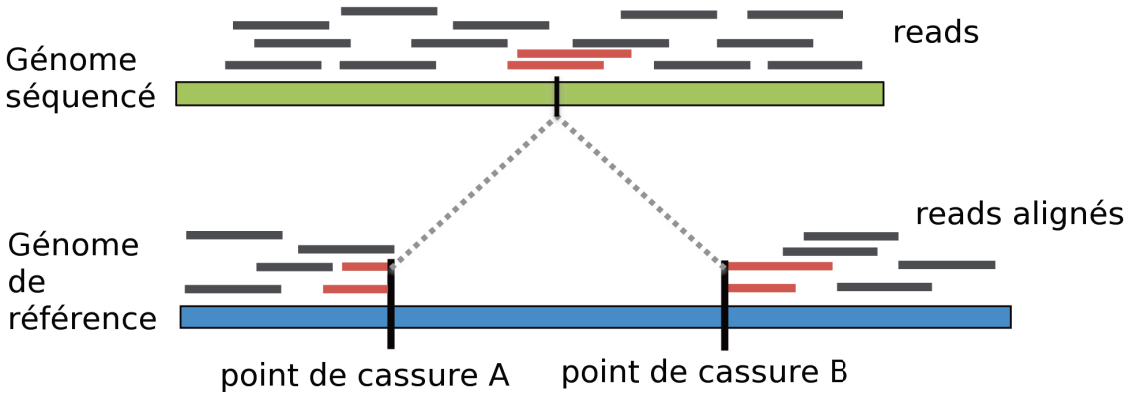
Figure 48 – Exemple de paires de reads concordante et discordante dans le cas d'une large délétion chromosomique. *adapté d'après Raphael et al. 2012*[325]



La section rouge représente une portion génomique perdue, et les segments verticaux noirs correspondent aux points de cassures définissant cette délétion.

Ainsi, la détection à large échelle des paires discordantes permet d'identifier les variants structuraux potentiellement présent au sein d'un génome tumoral. Une autre technique similaire consiste à exploiter une catégorie de reads dont la particularité est qu'ils *clippés*, c'est à dire qu'une portion située à une extrémité du read ne s'aligne pas dans la continuité du reste du fragment. Cette partie est dite *clippée*. En effet, si un read contient un des points de cassures (une des extrémités de la portion réarrangée), seule une portion du read s'alignera correctement sur le génome de référence. Certaines méthodes d'analyse telles que CREST[326] utilisent les reads *clippés* afin d'identifier les points de cassures potentiels et de reconstruire l'architecture du génome (Figure 49).

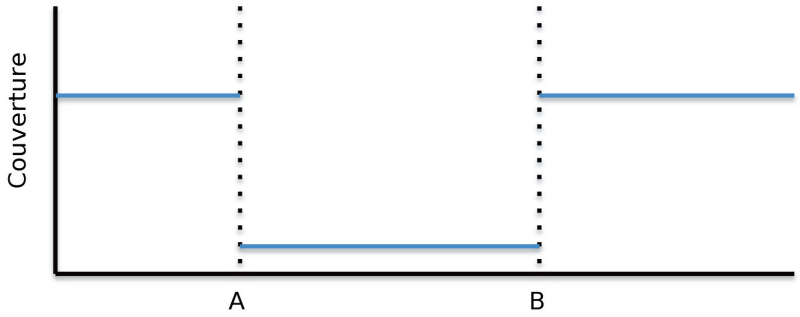
Figure 49 – Définition des points de cassures et reconstitution des variants structuraux à partir de reads *clippés. adapté d'après Raphael et al. 2012* [325]



Les segments noirs verticaux définissent les points de cassures, extrémités du variant structural représenté, qui dans cet exemple, est une délétion. Les reads représentés en rouge sont *clippés* : ils s'alignent sur le génome de référence en deux contigs distincts situés de part et d'autre des points de cassures.

Un troisième ensemble de méthodes se focalise sur l'étude de la couverture afin de détecter les variants structuraux. En effet, les portions chromosomiques qui subissent des réarrangements structuraux sont souvent également sujettes aux changements de nombre de copies. Cela se traduit par un changement brusque de la couverture de long du génome. Ces méthodes essayent donc de modéliser la couverture le long du génome afin de détecter les points de cassures (Figure 50). Ces méthodes sont donc capables de détecter uniquement des réarrangements caractérisés par une changement du nombre de copie d'une portion chromosomique. Elles pourront détecter des délétions ou des duplications, mais pas d'inversions. Le temps d'exécution de ce genre d'algorithme est cependant extrêmement long et requiert d'importantes ressources computationnelles. De plus, les données de couverture présente une variabilité plus ou moins forte en fonction de la plateforme de séquençage, et peuvent contenir beaucoup de bruit.

Figure 50 – Localisation des points de cassures par analyse de la couverture le long du génome. *adapté d'après Raphael et al. 2012* [325]



Représentation schématique de la couverture le long du génome. La courbe bleue représente la couverture. Les pointillés représentent les points de cassures A et B, entre lesquels la couverture chute.

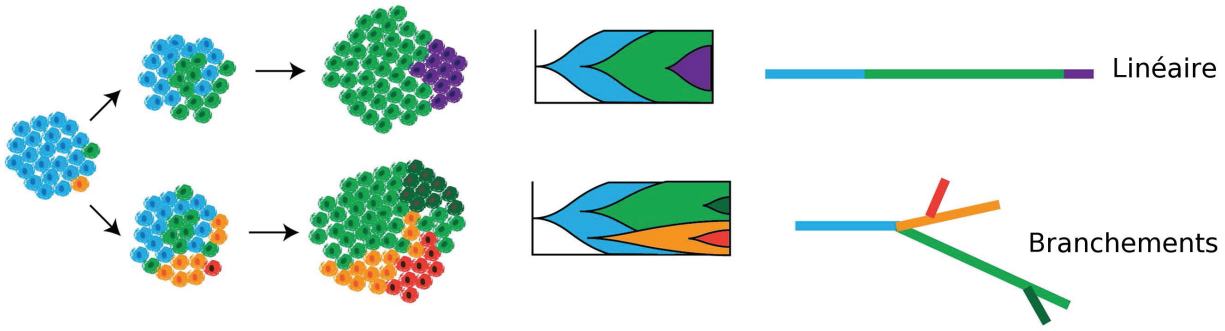
Enfin, la dernière méthode repose sur l'assemblage *de novo* des reads, c'est à dire la reconstitution du génome étudié, mais sans tenir compte du génome de référence. L'idée est qu'en s'affranchissant de la contrainte que l'on impose en alignant contre le génome de référence, on peut plus facilement reconstituer le génome étudié tel qu'il est réellement. Cependant, l'assemblage *de novo* d'un génome d'une taille aussi importante que celle du génome humain est extrêmement long, requiers également de très lourdes ressources computationnelles, mais est surtout d'une très forte complexité algorithmique.

Malgré l'existence des différents types de méthodes présentés ci-dessus, les performances de la communauté scientifique dans l'identification de variants structuraux restent limitées. De nombreuses équipes de recherche travaillent sur cette problématique à l'échelle internationale, mais les résultats ne sont pour l'instant pas à la hauteur de l'investissement fourni. En effet, il apparaît que la détection des variants structuraux est un problème beaucoup plus complexe qu'anticipé. C'est pourquoi, dans l'objectif de mutualiser les compétences des différents acteurs internationaux travaillant sur ce thème, l'ICGC - *International Cancer Genome Consortium* - et le TCGA -*The Cancer Genome Atlas*- ont lancé un projet intitulé « Challenge d'appel de variants somatiques », dont le but est de comparer les performances de différentes méthodes d'appel de variants somatiques, incluant les variants structuraux. Les plateformes de recherche bioinformatique participant à ce challenge ont donc toutes travaillé sur un jeu de données simulées commun. Les résultats préliminaires de ce challenge indiquent que parmi les variants structuraux identifiés, seuls 30% se recoupent sur l'ensemble des méthodes de prédiction. Il y a donc encore un très gros travail à réaliser afin d'optimiser les performances de détection de ce type d'altérations.

Une autre problématique émergente en bioinformatique du génome du cancer est la détection des sous-clones d'une tumeur. En effet, le génome de la tumeur évolue au cours de son développement, et suite à l'apparition de mutations et à la division de la cellule dans laquelle ces mutations sont apparues, une partie de la tumeur présentera une génome légèrement différent de la cellule tumorale initiale. Une tumeur peut ainsi être constituée par un ensemble de

cellules formant plusieurs sous-clones, c'est à dire plusieurs groupes de cellules ayant un génome propre. Alors que les mutations apparues tôt dans le développement tumoral seront présentes dans la majorité des sous-clones, les plus récentes ne seront présentes que dans un ou deux sous-clones. À l'heure actuelle différents modèles sont envisagés afin de comprendre l'histoire de développement d'une tumeur (Figure 51)

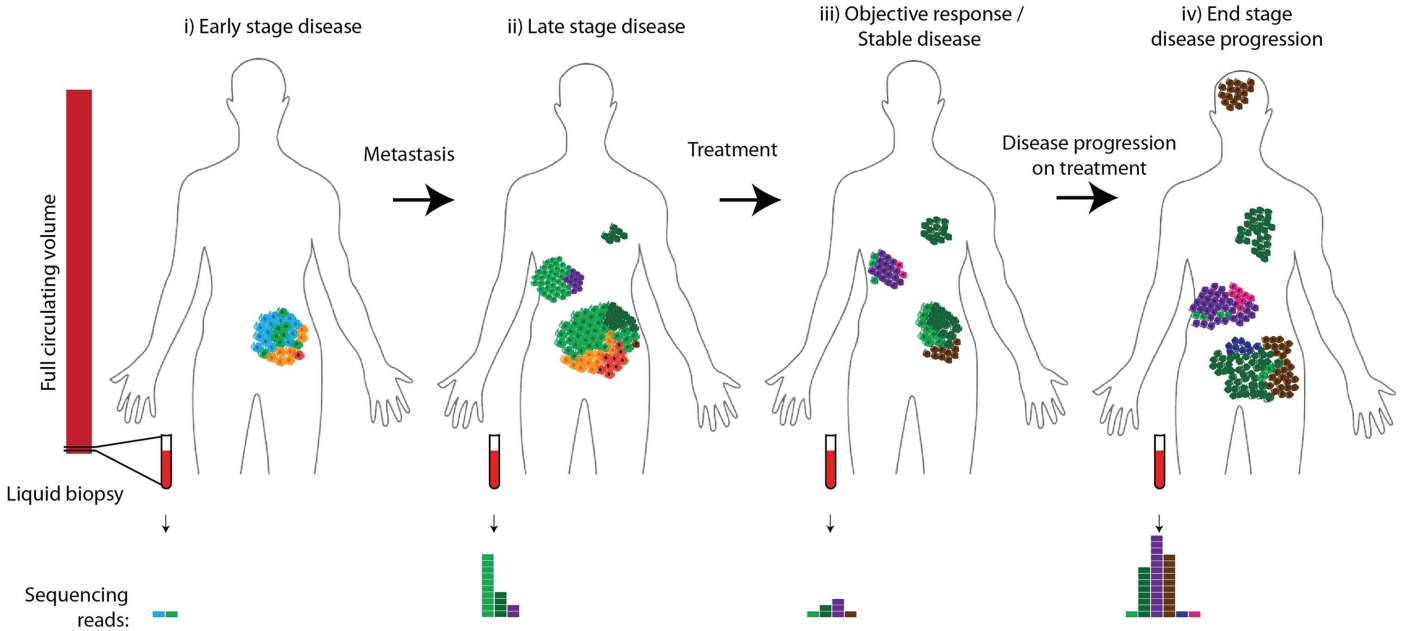
Figure 51 – Différents modèles pour l'évolution polyclonale de la tumeur *adapté d'après Burrell et al. 2014* [327]



La caractérisation des sous-clones tumoraux nécessite de pouvoir identifier quelles sont les spécificités génomiques de chacun d'entre eux, c'est à dire identifier quelles mutations les caractérisent. Or, l'ADN tumoral extrait d'une biopsie de la tumeur peut provenir de cellules appartenant à différents sous-clones, on séquence donc de l'ADN correspondant à un mélange de plusieurs génomes tumoraux possédant quelques différences les uns par rapport aux autres. De plus, une biopsie n'est qu'un prélèvement partiel de la tumeur. Dans un génome normal, les mutations observées sont soit homozygotes, soit hétérozygotes, donnant ainsi des fréquences attendues de 1 ou 0.5. Or, entre les variations de ploïdie et l'existence de sous-clones au sein du génome tumoral, le spectre des fréquences des variants observés change complètement, et les analyses de données de séquençage mettent en évidence de nombreux variants de fréquence variable entre 0 et 1, et notamment de faible fréquence. Une partie de ceux-ci sont très probablement des faux-positifs, mais une autre peut tout à fait correspondre à des variants spécifiques de sous-clones représentant une fraction minoritaire de la tumeur.

La tumeur peut être comparée à un écosystème dans lequel chaque cellule tumorale constitue une individualité ayant certaines capacités et soumise à un environnement donné. Les cellules tumorales constituent donc une population d'individus soumise à des pressions de sélection exercées par leur environnement. Une forme de sélection naturelle s'applique donc sur les cellules tumorales, et les cellules des sous-clones possédant des caractéristiques leur permettant de mieux résister à ces pressions prolifèrent. Les cellules d'autres sous-clones ne survivront pas face à ces pressions. La conséquence de ce phénomène est l'apparition de la résistance polyclonale au cours du traitement (Figure 52).

Figure 52 – Résistance polyclonale
d'après Burrell et al. 2014 [327]



En effet, les cellules de certains sous-clones tumoraux vont être sensibles aux traitements administrés. Cependant, il est possible que d'autres se montrent résistants à un traitement donné grâce aux spécificités génomiques qu'elles possèdent. Ces cellules vont alors proliférer, former des métastases qui migrent dans d'autres organes. On assiste ainsi tout d'abord à une réduction du volume tumoral, puis la maladie recommence à progresser même sous traitement. Cette résistance empêche de freiner la progression de la maladie et la colonisation d'autres organes, ce qui est souvent fatal. Il est donc d'autant plus important de détecter le plus tôt possible les mutations sous-clonales afin de pouvoir réagir en ajustant le traitement en cours de prise en charge.

La bioinformatique et les biostatistiques font également face à des écueils méthodologiques en épidémiologie génétique. Depuis une dizaine d'années, les études pangénomiques ont permis d'identifier dans notre génome constitutionnel nombre de loci de prédisposition à toute une variété de pathologies ayant une composante d'origine génétique. Cependant, le design des études ainsi que nos modèles statistiques actuels ne permettent de détecter que les polymorphismes ayant individuellement un effet mesurable au-dessus d'un certain seuil, déterminé par les caractéristiques de l'étude. Dans les analyses de données de GWAS, on espère trouver des SNPs dont la p-value d'association avec le risque de maladie estimé par régression logistique soit inférieure à 10^{-8}. C'est parfois le cas, et si plusieurs de ces candidats identifiés appartiennent à un même bloc de déséquilibre de liaison, l'étude est un succès. Cependant, une puce GWAS contient plusieurs centaines de milliers jusqu'à plusieurs millions de SNPs. Dans les résultats des tests d'associations effectués se trouvent des centaines voire des milliers de SNPs dont la p-value d'association est intermédiaire : parmi ces SNPs, on est incapable de prédire lesquels

pourraient avoir un effet mineur. On est de même aujourd'hui incapable d'appréhender efficacement à large échelle l'effet conjoint de variations sur le risque de développer une pathologie donnée. Certaines approches ont commencé à émerger ces dernières années. Le principe commun de ces approches est, dans une perspective typiquement bayésienne, d'enrichir les données génotypiques par de l'information biologique[328–332]. En effet, on connaît de mieux en mieux nos gènes, mais surtout leurs fonctions, et leurs interactions. Ces interactions sont modélisées sous la forme de voies métaboliques et de voies de régulations. Le principe est le suivant : si on est incapable de détecter l'effet individuel d'un ensemble de polymorphismes, on pourrait peut-être le détecter en considérant l'effet conjoint de ces polymorphismes. Or, il est impossible de tester un à un l'intégralité des sous-ensembles de SNPs présents sur une puce de plusieurs centaines de milliers de loci. L'idée est d'utiliser les informations biologiques connues sur les réseaux de gènes afin de regrouper les polymorphismes, et ainsi de tester si leurs variations sont plus susceptibles d'altérer le fonctionnement de l'entité biologique qui les relie, que ce soit au niveau d'un exon, d'un gène ou d'une voie de régulation. Ce genre de démarche est d'autant plus prometteuse que son exactitude est amenée à augmenter au fil du développement de notre connaissance sur les réseaux d'interactions génomiques.

La bioinformatique, les biostatistiques et l'épidémiologie génétique peuvent être rassemblées sous le patronyme de bioanalyse, bien que ce vaste domaine ne se limite pas à ces trois champs disciplinaires. Quelque soit la spécialité des sciences de la vie évoquée, la bioanalyse fait aujourd'hui partie intégrante de la recherche, et notamment de la recherche médicale. Les découvertes réalisées jusqu'ici sont majeures, notamment en cancérologie. De nombreux progrès restent à faire, notamment en développement méthodologique et pour ce qui relève de la transposition des résultats à la clinique. Cependant, à l'image du parcours déjà réalisé, les futures avancées en recherche génomique médicale ne se feront certainement pas sans un recours à ces disciplines.

Bibliographie

[1] R. Wakeford. The cancer epidemiology of radiation. *Oncogene*, 23(38) :6404–6428, 2004.

[2] J. M. Matés, J. A. Segura, F. J. Alonso, and J. Márquez. Roles of dioxins and heavy metals in cancer and neurological diseases using ros-mediated mechanisms. *Free Radical Biology and Medicine*, 49(9) :1328–1341, Nov. 2010.

[3] E. Y. H. P. Lee and W. J. Muller. Oncogenes and tumor suppressor genes. *Cold Spring Harbor perspectives in biology*, 2(10) :a003236, Oct. 2010. PMID : 20719876 PMCID : PMC2944361.

[4] D. A. Spandidos. Oncogenes and tumor suppressor genes as paradigms in oncogenesis. *Journal of B.U.ON. : official journal of the Balkan Union of Oncology*, 12 Suppl 1 :S9–12, Sept. 2007. PMID : 17935284.

[5] K. Suzuki and H. Matsubara. Recent advances in p53 research and cancer treatment. *BioMed Research International*, 2011, June 2011.

[6] R. Brosh and V. Rotter. When mutants gain new powers : news from the mutant p53 field. *Nature reviews. Cancer*, 9(10) :701–713, Oct. 2009. PMID : 19693097.

[7] F. Ricceri, G. Matullo, and P. Vineis. Is there evidence of involvement of dna repair polymorphisms in human cancer ? *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 736(1–2) :117–121, Aug. 2012.

[8] K. J. Ruddy and E. P. Winer. Male breast cancer : risk factors, biology, diagnosis, treatment, and survivorship. *Annals of Oncology*, 24(6) :1434–1443, Jan. 2013. PMID : 23425944.

[9] P. E. Goss, C. Reid, M. Pintilie, R. Lim, and N. Miller. Male breast carcinoma. *Cancer*, 85(3) :629–639, Feb. 1999.

[10] S. H. Giordano, D. S. Cohen, A. U. Buzdar, G. Perkins, and G. N. Hortobagyi. Breast carcinoma in men : a population-based study. *Cancer*, 101(1) :51–57, July 2004. PMID : 15221988.

[11] J. Ferlay, I. Soerjomatam, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. Parkin, D. Forman, and F. Bray. Globocan 2012 v1.0, cancer incidence and mortality worldwide : Iarc cancerbase no. 11, 2013.

[12] C. DeSantis, J. Ma, L. Bryan, and A. Jemal. Breast cancer statistics, 2013. *CA : A Cancer Journal for Clinicians*, 64(1) :52–62, Jan. 2014.

[13] L. Liu, J. Zhang, A. H. Wu, M. C. Pike, and D. Deapen. Invasive breast cancer incidence trends by detailed race/ethnicity and age. *International Journal of Cancer*, 130(2) :395–404, Jan. 2012.

[14] R. G. Ziegler, R. N. Hoover, M. C. Pike, A. Hildesheim, A. M. Y. Nomura, D. W. West, A. H. Wu-Williams, L. N. Kolonel, P. L. Horn-Ross, J. F. Rosenthal, and M. B. Hyer. Migration patterns and breast cancer risk in asian-american women. *Journal of the National Cancer Institute*, 85(22) :1819–1827, Nov. 1993.

[15] P. Eroles, A. Bosch, J. A. Pérez-Fidalgo, and A. Lluch. Molecular biology in breast cancer : intrinsic subtypes and signaling pathways. *Cancer treatment reviews*, 38(6) :698–707, Oct. 2012. PMID : 22178455.

[16] D. J. Slamon, B. Leyland-Jones, S. Shak, H. Fuchs, V. Paton, A. Bajamonde, T. Fleming, W. Eiermann, J. Wolter, M. Pegram, J. Baselga, and L. Norton. Use of chemotherapy plus a monoclonal antibody against her2 for metastatic breast cancer that overexpresses her2. *The New England journal of medicine*, 344(11) :783–792, Mar. 2001. PMID : 11248153.

[17] C. A. Hudis. Trastuzumab —mechanism of action and use in clinical practice. *New England Journal of Medicine*, 357(1) :39–51, 2007. PMID : 17611206.

[18] J. S. Ross, E. A. Slodkowska, W. F. Symmans, L. Pusztai, P. M. Ravdin, and G. N. Hortobagyi. The her-2 receptor and breast cancer : ten years of targeted anti-her-2 therapy and personalized medicine. *The oncologist*, 14(4) :320–368, Apr. 2009. PMID : 19346299.

[19] N. U. Lin, A. Vanderplas, M. E. Hughes, R. L. Theriault, S. B. Edge, Y.-N. Wong, D. W. Blayney, J. C. Niland, E. P. Winer, and J. C. Weeks. Clinicopathologic features, patterns of recurrence, and survival among women with triple-negative breast cancer in the national comprehensive cancer network. *Cancer*, 118(22) :5463–5472, Nov. 2012. PMID : 22544643 PMCID : PMC3611659.

[20] R. Dent, M. Trudeau, K. I. Pritchard, W. M. Hanna, H. K. Kahn, C. A. Sawka, L. A. Lickley, E. Rawlinson, P. Sun, and S. A. Narod. Triple-negative breast cancer : clinical features and patterns of recurrence. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 13(15 Pt 1) :4429–4434, Aug. 2007. PMID : 17671126.

[21] P. G. Morris, C. G. Murphy, D. Mallam, M. Accordino, S. Patil, J. Howard, A. Omuro, K. Beal, A. D. Seidman, C. A. Hudis, and M. N. Fornier. Limited overall survival in patients with brain metastases from triple negative breast cancer. *The breast journal*, 18(4) :345–350, Aug. 2012. PMID : 22607041.

[22] A. Prat, B. Adamo, M. C. U. Cheang, C. K. Anders, L. A. Carey, and C. M. Perou. Molecular characterization of basal-like and non-basal-like triple-negative breast cancer. *The oncologist*, 18(2) :123–133, 2013. PMID : 23404817 PMCID : PMC3579595.

[23] M. J. Kwon. Emerging roles of claudins in human cancer. *International journal of molecular sciences*, 14(9) :18148–18180, 2013. PMID : 24009024 PMCID : PMC3794774.

[24] A. Tessari, D. Palmieri, and S. Di Cosimo. Overview of diagnostic/targeted treatment combinations in personalized medicine for breast cancer patients. *Pharmacogenomics and personalized medicine*, 7 :1–19, 2013. PMID : 24403841 PMCID : PMC3883531.

[25] S. Boyault, Y. Drouet, C. Navarro, T. Bachelot, C. Lasset, I. Treilleux, E. Tabone, A. Puisieux, and Q. Wang. Mutational characterization of individual breast tumors : Tp53 and pi3k pathway genes are frequently and distinctively mutated in different subtypes. *Breast cancer research and treatment*, 132(1) :29–39, Feb. 2012. PMID : 21512767.

[26] M. Tenhagen, P. J. van Diest, I. A. Ivanova, E. van der Wall, and P. van der Groep. Fibroblast growth factor receptors in breast cancer : expression, downstream effects, and possible drug targets. *Endocrine-related cancer*, 19(4) :R115–129, Aug. 2012. PMID : 22508544.

[27] L. Melchor and J. Benítez. The complex genetic landscape of familial breast cancer. *Human genetics*, 132(8) :845–863, Aug. 2013. PMID : 23552954.

[28] M. Ghoussaini, P. D. P. Pharoah, and D. F. Easton. Inherited genetic susceptibility to breast cancer : The beginning of the end or the end of the beginning ? *The American Journal of Pathology*, 183(4) :1038–1051, Oct. 2013.

[29] A. M. Martin, M. A. Blackwood, D. Antin-Ozerkis, H. A. Shih, K. Calzone, T. A. Colligon, S. Seal, N. Collins, M. R. Stratton, B. L. Weber, and K. L. Nathanson. Germline mutations in brca1 and brca2 in breast-ovarian families from a breast cancer risk evaluation clinic. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, 19(8) :2247–2253, Apr. 2001. PMID : 11304778.

[30] D. G. Evans, A. Shenton, E. Woodward, F. Lalloo, A. Howell, and E. R. Maher. Penetrance estimates for brca1 and brca2 based on genetic testing in a clinical cancer genetics service setting : risks of breast/ovarian cancer quoted should reflect the cancer burden in the family. *BMC cancer*, 8 :155, 2008. PMID : 18513387.

[31] A. Antoniou, P. D. P. Pharoah, S. Narod, H. A. Risch, J. E. Eyfjord, J. L. Hopper, N. Loman, H. Olsson, O. Johannsson, A. Borg, B. Pasini, P. Radice, S. Manoukian, D. M. Eccles, N. Tang, E. Olah, H. Anton-Culver, E. Warner, J. Lubinski, J. Gronwald, B. Gorski, H. Tulinius, S. Thorlacius, H. Eerola, H. Nevanlinna, K. Syrjäkoski, O.-P. Kallioniemi, D. Thompson, C. Evans, J. Peto, F. Lalloo, D. G. Evans, and D. F. Easton. Average risks of breast and ovarian cancer associated with brca1 or brca2 mutations detected in case series unselected for family history : a combined analysis of 22 studies. *American journal of human genetics*, 72(5) :1117–1130, May 2003. PMID : 12677558.

[32] J. P. Struewing, D. Abeliovich, T. Peretz, N. Avishai, M. M. Kaback, F. S. Collins, and L. C. Brody. The carrier frequency of the brca1 185delag mutation is approximately 1 percent in ashkenazi jewish individuals. *Nature Genetics*, 11(2) :198–200, Oct. 1995.

[33] C. Oddoux, J. P. Struewing, C. M. Clayton, S. Neuhausen, L. C. Brody, M. Kaback, B. Haas, L. Norton, P. Borgen, S. Jhanwar, D. Goldgar, H. Ostrer, and K. Offit. The carrier frequency of the brca2 6174delt mutation among ashkenazi jewish individuals is approximately 1%. *Nature Genetics*, 14(2) :188–190, Oct. 1996.

[34] P. Hartge, J. P. Struewing, S. Wacholder, L. C. Brody, and M. A. Tucker. The prevalence of common brca1 and brca2 mutations among ashkenazi jews. *American Journal of Human Genetics*, 64(4) :963–970, Apr. 1999. PMID : 10090881 PMCID : PMC1377820.

[35] R. Ferla, V. Calò, S. Cascio, G. Rinaldi, G. Badalamenti, I. Carreca, E. Surmacz, G. Colucci, V. Bazan, and A. Russo. Founder mutations in brca1 and brca2 genes. *Annals of Oncology*, 18(suppl 6) :vi93–vi98, June 2007. PMID : 17591843.

[36] D. Malkin, F. P. Li, L. C. Strong, J. Fraumeni, J F, C. E. Nelson, D. H. Kim, J. Kassel, M. A. Gryka, F. Z. Bischoff, and M. A. Tainsky. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science (New York, N.Y.)*, 250(4985) :1233–1238, Nov. 1990. PMID : 1978757.

[37] S. Srivastava, Z. Q. Zou, K. Pirollo, W. Blattner, and E. H. Chang. Germ-line transmission of a mutated p53 gene in a cancer-prone family with li-fraumeni syndrome. *Nature*, 348(6303) :747–749, Dec. 1990. PMID : 2259385.

[38] J. M. Birch, A. L. Hartley, K. J. Tricker, J. Prosser, A. Condie, A. M. Kelsey, M. Harris, P. H. Jones, A. Binchy, and D. Crowther. Prevalence and diversity of constitutional mutations in the p53 gene among 21 li-fraumeni families. *Cancer research*, 54(5) :1298–1304, Mar. 1994. PMID : 8118819.

[39] K. E. Nichols, D. Malkin, J. E. Garber, J. F. Fraumeni, and F. P. Li. Germ-line p53 mutations predispose to a wide spectrum of early-onset cancers. *Cancer Epidemiology Biomarkers & Prevention*, 10(2) :83–87, Jan. 2001. PMID : 11219776.

[40] P. A. J. Muller and K. H. Vousden. p53 mutations in cancer. *Nature cell biology*, 15(1) :2–8, Jan. 2013. PMID : 23263379.

[41] N. Rahman, S. Seal, D. Thompson, P. Kelly, A. Renwick, A. Elliott, S. Reid, K. Spanova, R. Barfoot, T. Chagtai, H. Jayatilake, L. McGuffog, S. Hanks, D. G. Evans, D. Eccles, D. F. Easton, and M. R. Stratton. Palb2, which encodes a brca2-interacting protein, is a breast cancer susceptibility gene. *Nature Genetics*, 39(2) :165–167, Feb. 2007.

[42] H. Meijers-Heijboer, A. v. d. Ouweland, J. Klijn, M. Wasielewski, A. d. Snoo, R. Oldenburg, A. Hollestelle, M. Houben, E. Crepin, M. v. Veghel-Plandsoen, F. Elstrodt, C. v. Duijn, C. Bartels, C. Meijers, M. Schutte, L. McGuffog, D. Thompson, D. F. Easton, N. Sodha, S. Seal, R. Barfoot, J. Mangion, J. Chang-Claude, D. Eccles, R. Eeles, D. G. Evans, R. Houlston, V. Murday, S. Narod, T. Peretz, J. Peto, C. Phelan, H. X. Zhang, C. Szabo, P. Devilee, D. Goldgar, P. A. Futreal, K. L. Nathanson, B. L. Weber, N. Rahman, and M. R. Stratton. Low-penetrance susceptibility to breast cancer due to chek2∗1100delc in noncarriers of brca1 or brca2 mutations. *Nature Genetics*, 31(1) :55–59, May 2002.

[43] A. Renwick, D. Thompson, S. Seal, P. Kelly, T. Chagtai, M. Ahmed, B. North, H. Jayatilake, R. Barfoot, K. Spanova, L. McGuffog, D. G. Evans, D. Eccles, D. F. Easton, M. R. Stratton, and N. Rahman. Atm mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nature Genetics*, 38(8) :873–875, Aug. 2006.

[44] S. Seal, D. Thompson, A. Renwick, A. Elliott, P. Kelly, R. Barfoot, T. Chagtai, H. Jayatilake, M. Ahmed, K. Spanova, B. North, L. McGuffog, D. G. Evans, D. Eccles, D. F. Easton, M. R. Stratton, and N. Rahman. Truncating mutations in the fanconi anemia j gene brip1 are low-penetrance breast cancer susceptibility alleles. *Nature Genetics*, 38(11) :1239–1241, Nov. 2006.

[45] A. De Nicolo, M. Tancredi, G. Lombardi, C. C. Flemma, S. Barbuti, C. Di Cristofano, B. Sobhian, G. Bevilacqua, R. Drapkin, and M. A. Caligo. A novel breast cancer-associated brip1 (fancj/bach1) germ-line mutation impairs protein stability and function. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 14(14) :4672–4680, July 2008. PMID : 18628483 PMCID : PMC2561321.

[46] N. Resta, D. Pierannunzio, G. M. Lenato, A. Stella, R. Capocaccia, R. Bagnulo, P. Lastella, F. C. Susca, C. Bozzao, D. C. Loconte, C. Sabbà, E. Urso, P. Sala, M. Fornasarig, P. Grammatico, A. Piepoli, C. Host, D. Turchetti, A. Viel, L. Memo, L. Giunti, V. Stigliano, L. Varesco, L. Bertario, M. Genuardi, E. Lucci Cordisco, M. G. Tibiletti, C. Di Gregorio, A. Andriulli, and M. Ponz de Leon. Cancer risk associated with stk11/lkb1 germline mutations in peutz–jeghers syndrome patients : Results of an italian multicenter study. *Digestive and Liver Disease*, 45(7) :606–611, July 2013.

[47] F. M. Giardiello, J. D. Brensinger, A. C. Tersmette, S. N. Goodman, G. M. Petersen, S. V. Booker, M. Cruz-Correa, and J. A. Offerhaus. Very high risk of cancer in familial peutz-jeghers syndrome. *Gastroenterology*, 119(6) :1447–1453, Dec. 2000. PMID : 11113065.

[48] P. Guilford, J. Hopkins, J. Harraway, M. McLeod, N. McLeod, P. Harawira, H. Taite, R. Scoular, A. Miller, and A. E. Reeve. E-cadherin germline mutations in familial gastric cancer. *Nature*, 392(6674) :402–405, Mar. 1998.

[49] C. Petridis, I. Shinomiya, K. Kohut, P. Gorman, M. Caneppele, V. Shah, M. Troy, S. E. Pinder, A. Hanby, I. Tomlinson, R. C. Trembath, R. Roylance, M. A. Simpson, and E. J. Sawyer. Germline cdh1 mutations in bilateral lobular carcinoma in situ. *British Journal of Cancer*, 110(4) :1053–1057, Feb. 2014.

[50] N. Bogdanova, S. Helbig, and T. Dörk. Hereditary breast cancer : ever more pieces to the polygenic puzzle. *Hereditary cancer in clinical practice*, 11(1) :12, 2013. PMID : 24025454 PMCID : PMC3851033.

[51] D. C. Wallace. A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer : A dawn for evolutionary medicine. *Annual Review of Genetics*, 39(1) :359–407, Dec. 2005.

[52] C. A. Clarke, T. H. M. Keegan, J. Yang, D. J. Press, A. W. Kurian, A. H. Patel, and J. Lacey, James V. Age-specific incidence of breast cancer subtypes : understanding the black-white crossover. *Journal of the National Cancer Institute*, 104(14) :1094–1101, July 2012. PMID : 22773826 PMCID : PMC3640371.

[53] P. Irigaray, J. A. Newby, R. Clapp, L. Hardell, V. Howard, L. Montagnier, S. Epstein, and D. Belpomme. Lifestyle-related factors and environmental agents causing cancer : An overview. *Biomedicine & Pharmacotherapy*, 61(10) :640–658, Dec. 2007.

[54] C. A. Thomson, M. L. McCullough, B. C. Wertheim, R. T. Chlebowski, M. E. Martinez, M. L. Stefanick, T. E. Rohan, J. E. Manson, H. A. Tindle, J. Ockene, M. Z. Vitolins, J. Wactawski-Wende, G. E. Sarto, D. S. Lane, and M. L. Neuhouser. Nutrition and physical activity cancer prevention guidelines, cancer risk, and mortality in the women's health initiative. *Cancer prevention research (Philadelphia, Pa.)*, 7(1) :42–53, Jan. 2014. PMID : 24403289.

[55] T. L. L. Thomas P Ahern. Lifetime tobacco smoke exposure and breast cancer incidence. *Cancer causes &amp ; control : CCC*, 20(10) :1837–44, 2009.

[56] T. L. Lash and A. Aschengrau. A null association between active or passive cigarette smoking and breast cancer risk. *Breast Cancer Research and Treatment*, 75(2) :181–184, Sept. 2002. PMID : 12243511.

[57] Y. Lin, S. Kikuchi, K. Tamakoshi, K. Wakai, T. Kondo, Y. Niwa, H. Yatsuya, K. Nishio, S. Suzuki, S. Tokudome, A. Yamamoto, H. Toyoshima, M. Mori, A. Tamakoshi, and Japan Collaborative Cohort Study Group for Evaluation of Cancer Risk. Active smoking, passive smoking, and breast cancer risk : findings from the japan collaborative cohort study for evaluation of cancer risk. *Journal of Epidemiology / Japan Epidemiological Association*, 18(2) :77–83, 2008. PMID : 18403857.

[58] J. Prescott, H. Ma, L. Bernstein, and G. Ursin. Cigarette smoking is not associated with breast cancer risk in young women. *Cancer Epidemiology Biomarkers & Prevention*, 16(3) :620–622, Jan. 2007. PMID : 17372262.

[59] N. A. Field, M. S. Baptiste, P. C. Nasca, and B. B. Metzger. Cigarette smoking and breast cancer. *International Journal of Epidemiology*, 21(5) :842–848, Oct. 1992. PMID : 1468843.

[60] N. Hamajima, K. Hirose, K. Tajima, T. Rohan, E. E. Calle, C. W. Heath, R. J. Coates, J. M. Liff, R. Talamini, N. Chantarakul, S. Koetsawang, D. Rachawat, A. Morabia, L. Schuman, W. Stewart, M. Szklo, C. Bain, F. Schofield, V. Siskind, P. Band, A. J. Coldman, R. P. Gallagher, T. G. Hislop, P. Yang, L. M. Kolonel, A. M. Y. Nomura, J. Hu, K. C. Johnson, Y. Mao, S. De Sanjosé, N. Lee, P. Marchbanks, H. W. Ory, H. B. Peterson, H. G. Wilson, P. A. Wingo, K. Ebeling, D. Kunde, P. Nishan, J. L. Hopper, G. Colditz, V. Gajalanski, N. Martin, T. Pardthaisong, S. Silpisornkosol, C. Theetranont, B. Boosiri, S. Chutivongse, P. Jimakorn, P. Virutamasen, C. Wongsrichanalai, M. Ewertz, H. O. Adami, L. Bergkvist, C. Magnusson, I. Persson, J. Chang-Claude, C. Paul, D. C. G. Skegg, G. F. S. Spears, P. Boyle, T. Evstifeeva, J. R. Daling, W. B. Hutchinson, K. Malone, E. A. Noonan, J. L. Stanford, D. B. Thomas, N. S. Weiss, E. White, N. Andrieu, A. Brêmond, F. Clavel, B. Gairard, J. Lansac, L. Piana, R. Renaud, A. Izquierdo, P. Viladiu, H. R. Cuevas, P. Ontiveros, A. Palet, S. B. Salazar, N. Aristizabel, A. Cuadros, L. Tryggvadottir, H. Tulinius, A. Bachelot, M. G. Lê, J. Peto, S. Franceschi, F. Lubin, B. Modan, E. Ron, Y. Wax, G. D. Friedman, R. A. Hiatt, F. Levi, T. Bishop, K. Kosmelj, M. Primic-Zakelj, B. Ravnihar, J. Stare, W. L. Beeson, G. Fraser, R. D. Bullbrook, J. Cuzick, S. W. Duffy, I. S. Fentiman, J. L. Hayward, D. Y. Wang, A. J. McMichael, K. McPherson, R. L. Hanson, M. C. Leske, M. C. Mahoney, P. C. Nasca, A. O. Varma,

A. L. Weinstein, T. R. Moller, H. Olsson, J. Ranstam, R. A. Goldbohm, P. A. van den Brandt, R. A. Apelo, J. Baens, J. R. de la Cruz, B. Javier, L. B. Lacaya, C. A. Ngelangel, C. La Vecchia, E. Negri, E. Marubini, M. Ferraroni, M. Gerber, S. Richardson, C. Segala, D. Gatei, P. Kenya, A. Kungu, J. G. Mati, L. A. Brinton, R. Hoover, C. Schairer, R. Spirtas, H. P. Lee, M. A. Rookus, F. E. van Leeuwen, J. A. Schoenberg, M. McCredie, M. D. Gammon, E. A. Clarke, L. Jones, A. Neil, M. Vessey, D. Yeates, P. Appleby, E. Banks, V. Beral, D. Bull, B. Crossley, A. Goodill, J. Green, C. Hermon, T. Key, N. Langston, C. Lewis, G. Reeves, R. Collins, R. Doll, R. Peto, K. Mabuchi, D. Preston, P. Hannaford, C. Kay, L. Rosero-Bixby, Y. T. Gao, F. Jin, J.-M. Yuan, H. Y. Wei, T. Yun, C. Zhiheng, G. Berry, J. Cooper Booth, T. Jelihovsky, R. MacLennan, R. Shearman, Q.-S. Wang, C.-J. Baines, A. B. Miller, C. Wall, E. Lund, H. Stalsberg, X. O. Shu, W. Zheng, K. Katsouyanni, A. Trichopoulou, D. Trichopoulos, A. Dabancens, L. Martinez, R. Molina, O. Salas, F. E. Alexander, K. Anderson, A. R. Folsom, B. S. Hulka, L. Bernstein, S. Enger, R. W. Haile, A. Paganini-Hill, M. C. Pike, R. K. Ross, G. Ursin, M. C. Yu, M. P. Longnecker, P. Newcomb, L. Bergkvist, A. Kalache, T. M. M. Farley, S. Holck, O. Meirik, and Collaborative Group on Hormonal Factors in Breast Cancer. Alcohol, tobacco and breast cancer–collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease. *British Journal of Cancer*, 87(11) :1234–1245, Nov. 2002. PMID : 12439712 PMCID : PMC2562507.

[61] K. C. Johnson, J. Hu, Y. Mao, and Canadian Cancer Registries Epidemiology Research Group. Passive and active smoking and breast cancer risk in canada, 1994-97. *Cancer causes & control : CCC*, 11(3) :211–221, Mar. 2000. PMID : 10782655.

[62] Y. Cui, A. B. Miller, and T. E. Rohan. Cigarette smoking and breast cancer risk : update of a prospective cohort study. *Breast Cancer Research and Treatment*, 100(3) :293–299, Dec. 2006. PMID : 16773435.

[63] I. T. Gram, T. Braaten, P. D. Terry, A. J. Sasco, H.-O. Adami, E. Lund, and E. Weiderpass. Breast cancer risk among women who start smoking as teenagers. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, 14(1) :61–66, Jan. 2005. PMID : 15668477.

[64] F. Xue, W. C. Willett, B. A. Rosner, S. E. Hankinson, and K. B. Michels. Cigarette smoking and the incidence of breast cancer. *Archives of Internal Medicine*, 171(2) :125–133, Jan. 2011. PMID : 21263102 PMCID : PMC3131146.

[65] M. Cotterchio, L. Mirea, H. Ozcelik, and N. Kreiger. Active cigarette smoking, variants in carcinogen metabolism genes and breast cancer risk among pre- and postmenopausal women in ontario, canada. *The Breast Journal*, July 2014. PMID : 25052559.

[66] G. Sadri and H. Mahjub. Passive or active smoking, which is more relevant to breast cancer. *Saudi Medical Journal*, 28(2) :254–258, Feb. 2007. PMID : 17268706.

[67] K. C. Johnson, A. B. Miller, N. E. Collishaw, J. R. Palmer, S. K. Hammond, A. G. Salmon, K. P. Cantor, M. D. Miller, N. F. Boyd, J. Millar, and F. Turcotte. Active

smoking and secondhand smoke increase breast cancer risk : the report of the canadian expert panel on tobacco smoke and breast cancer risk (2009). *Tobacco Control*, 20(1) :e2, Jan. 2011. PMID : 21148114.

[68] C. E. Land, J. D. Boice, R. E. Shore, J. E. Norman, and M. Tokunaga. Breast cancer risk from low-dose exposures to ionizing radiation : Results of parallel analysis of three exposed populations of women. *Journal of the National Cancer Institute*, 65(2) :353–376, Jan. 1980. PMID : 6931253.

[69] D. L. Preston, A. Mattsson, E. Holmberg, R. Shore, N. G. Hildreth, and J. D. B. Jr. Radiation effects on breast cancer risk : A pooled analysis of eight cohorts. http ://www.rrjournal.org/doi/abs/10.1667/0033-7587July 2009.

[70] D. Cibula, A. Gompel, A. O. Mueck, C. La Vecchia, P. C. Hannaford, S. O. Skouby, M. Zikan, and L. Dusek. Hormonal contraception and risk of cancer. *Human reproduction update*, 16(6) :631–650, Dec. 2010. PMID : 20543200.

[71] A. Gompel and R. J. Santen. Hormone therapy and breast cancer risk 10 years after the whi. *Climacteric*, 15(3) :241–249, May 2012.

[72] A. Kendall, E. J. Folkerd, and M. Dowsett. Influences on circulating oestrogens in postmenopausal women : relationship with breast cancer. *The Journal of steroid biochemistry and molecular biology*, 103(2) :99–109, Feb. 2007. PMID : 17088056.

[73] V. Assi, J. Warwick, J. Cuzick, and S. W. Duffy. Clinical and epidemiological issues in mammographic density. *Nature reviews. Clinical oncology*, 9(1) :33–40, Jan. 2012. PMID : 22143145.

[74] S. Kobayashi, H. Sugiura, Y. Ando, N. Shiraki, T. Yanagi, H. Yamashita, and T. Toyama. Reproductive history and breast cancer risk. *Breast cancer (Tokyo, Japan)*, 19(4) :302–308, Oct. 2012. PMID : 22711317 PMCID : PMC3479376.

[75] A. G. Renehan, M. Tyson, M. Egger, R. F. Heller, and M. Zwahlen. Body-mass index and incidence of cancer : a systematic review and meta-analysis of prospective observational studies. *Lancet*, 371(9612) :569–578, Feb. 2008. PMID : 18280327.

[76] P. Manders, A. Pijpe, M. J. Hooning, I. Kluijt, H. F. A. Vasen, N. Hoogerbrugge, C. J. van Asperen, H. Meijers-Heijboer, M. G. E. M. Ausems, T. A. van Os, E. B. Gomez-Garcia, R. M. Brohet, HEBON, F. E. van Leeuwen, and M. A. Rookus. Body weight and risk of breast cancer in brca1/2 mutation carriers. *Breast Cancer Research and Treatment*, 126(1) :193–202, Feb. 2011. PMID : 20730487.

[77] J. R. Palmer, L. L. Adams-Campbell, D. A. Boggs, L. A. Wise, and L. Rosenberg. A prospective study of body size and breast cancer in black women. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, 16(9) :1795–1802, Sept. 2007. PMID : 17855697.

[78] R. Suzuki, N. Orsini, S. Saji, T. J. Key, and A. Wolk. Body weight and incidence of breast cancer defined by estrogen and progesterone receptor status–a meta-analysis. *International Journal of Cancer. Journal International Du Cancer*, 124(3) :698–712, Feb. 2009. PMID : 18988226.

[79] P. A. v. d. Brandt, D. Spiegelman, S.-S. Yaun, H.-O. Adami, L. Beeson, A. R. Folsom, G. Fraser, R. A. Goldbohm, S. Graham, L. Kushi, J. R. Marshall, A. B. Miller, T. Rohan, S. A. Smith-Warner, F. E. Speizer, W. C. Willett, A. Wolk, and D. J. Hunter. Pooled analysis of prospective cohort studies on height, weight, and breast cancer risk. *American Journal of Epidemiology*, 152(6) :514–527, Sept. 2000. PMID : 10997541.

[80] S. M. Enger, R. K. Ross, A. Paganini-Hill, C. L. Carpenter, and L. Bernstein. Body size, physical activity, and breast cancer hormone receptor status : Results from two case-control studies. *Cancer Epidemiology Biomarkers & Prevention*, 9(7) :681–687, Jan. 2000. PMID : 10919738.

[81] P. Berstad, R. J. Coates, L. Bernstein, S. G. Folger, K. E. Malone, P. A. Marchbanks, L. K. Weiss, J. M. Liff, J. A. McDonald, B. L. Strom, M. S. Simon, D. Deapen, M. F. Press, R. T. Burkman, R. Spirtas, and G. Ursin. A case-control study of body mass index and breast cancer risk in white and african-american women. *Cancer Epidemiology Biomarkers & Prevention*, 19(6) :1532–1544, Jan. 2010. PMID : 20501755.

[82] J. A. Britton, M. D. Gammon, J. B. Schoenberg, J. L. Stanford, R. J. Coates, C. A. Swanson, N. Potischman, K. E. Malone, D. J. Brogan, J. R. Daling, and L. A. Brinton. Risk of breast cancer classified by joint estrogen receptor and progesterone receptor status among women 20–44 years of age. *American Journal of Epidemiology*, 156(6) :507–516, Sept. 2002. PMID : 12225998.

[83] H. R. Harris, W. C. Willett, K. L. Terry, and K. B. Michels. Body fat distribution and risk of premenopausal breast cancer in the nurses' health study ii. *Journal of the National Cancer Institute*, 103(3) :273–278, Feb. 2011. PMID : 21163903.

[84] M. Harvie, L. Hooper, and A. Howell. Central obesity and breast cancer risk : a systematic review. *Obesity Reviews*, 4(3) :157–173, Aug. 2003.

[85] W.-Y. Huang, B. Newman, R. C. Millikan, M. J. Schell, B. S. Hulka, and P. G. Moorman. Hormone-related factors and risk of breast cancer in relation to estrogen receptor and progesterone receptor status. *American Journal of Epidemiology*, 151(7) :703–714, Jan. 2000. PMID : 10752798.

[86] P. H. Lahmann, K. Hoffmann, N. Allen, C. H. van Gils, K.-T. Khaw, B. Tehard, F. Berrino, A. Tjønneland, J. Bigaard, A. Olsen, K. Overvad, F. Clavel-Chapelon, G. Nagel, H. Boeing, D. Trichopoulos, G. Economou, G. Bellos, D. Palli, R. Tumino, S. Panico, C. Sacerdote, V. Krogh, P. H. M. Peeters, H. B. Bueno-de Mesquita, E. Lund, E. Ardanaz, P. Amiano, G. Pera, J. R. Quirós, C. Martínez, M. J. Tormo, E. Wirfält, G. Berglund, G. Hallmans, T. J. Key, G. Reeves, S. Bingham, T. Norat, C. Biessy, R. Kaaks, and E. Riboli. Body size and breast cancer risk : findings from the european prospective investigation into cancer and nutrition (epic). *International Journal of Cancer. Journal International Du Cancer*, 111(5) :762–771, Sept. 2004. PMID : 15252848.

[87] T. O. Ogundiran, D. Huo, A. Adenipekun, O. Campbell, R. Oyesegun, E. Akang, C. Adebamowo, and O. I. Olopade. Case-control study of body size and breast cancer risk in nigerian women. *American Journal of Epidemiology*, 172(6) :682–690, Sept. 2010. PMID : 20716701 PMCID : PMC2950817.

[88] A. Amadou, P. Ferrari, R. Muwonge, A. Moskal, C. Biessy, I. Romieu, and P. Hainaut. Overweight, obesity and risk of premenopausal breast cancer according to ethnicity : a systematic review and dose-response meta-analysis. *Obesity Reviews : An Official Journal of the International Association for the Study of Obesity*, 14(8) :665–678, Aug. 2013. PMID : 23615120.

[89] P. Boffetta and M. Hashibe. Alcohol and cancer. *The lancet oncology*, 7(2) :149–156, Feb. 2006. PMID : 16455479.

[90] G. D. Coronado, J. Beasley, and J. Livaudais. Alcohol consumption and the risk of breast cancer. *Salud pública de México*, 53(5) :440–447, Oct. 2011. PMID : 22218798.

[91] S. A. Qureshi, A. C. Lund, M. B. Veierød, M. H. Carlsen, R. Blomhoff, L. F. Andersen, and G. Ursin. Food items contributing most to variation in antioxidant intake ; a cross-sectional study among norwegian women. *BMC public health*, 14(1) :45, 2014. PMID : 24433390 PMCID : PMC3902183.

[92] S. Nickels, T. Truong, R. Hein, K. Stevens, K. Buck, S. Behrens, U. Eilber, M. Schmidt, L. Häberle, A. Vrieling, M. Gaudet, J. Figueroa, N. Schoof, A. B. Spurdle, A. Rudolph, P. A. Fasching, J. L. Hopper, E. Makalic, D. F. Schmidt, M. C. Southey, M. W. Beckmann, A. B. Ekici, O. Fletcher, L. Gibson, I. dos Santos Silva, J. Peto, M. K. Humphreys, J. Wang, E. Cordina-Duverger, F. Menegaux, B. G. Nordestgaard, S. E. Bojesen, C. Lanng, H. Anton-Culver, A. Ziogas, L. Bernstein, C. A. Clarke, H. Brenner, H. Müller, V. Arndt, C. Stegmaier, H. Brauch, T. Brüning, V. Harth, The GENICA Network, A. Mannermaa, V. Kataja, V.-M. Kosma, J. M. Hartikainen, kConFab, A. M. Group, D. Lambrechts, D. Smeets, P. Neven, R. Paridaens, D. Flesch-Janys, N. Obi, S. Wang-Gohrke, F. J. Couch, J. E. Olson, C. M. Vachon, G. G. Giles, G. Severi, L. Baglietto, K. Offit, E. M. John, A. Miron, I. L. Andrulis, J. A. Knight, G. Glendon, A. M. Mulligan, S. J. Chanock, J. Lissowska, J. Liu, A. Cox, H. Cramp, D. Connley, S. Balasubramanian, A. M. Dunning, M. Shah, A. Trentham-Dietz, P. Newcomb, L. Titus, K. Egan, E. K. Cahoon, P. Rajaraman, A. J. Sigurdson, M. M. Doody, P. Guénel, P. D. P. Pharoah, M. K. Schmidt, P. Hall, D. F. Easton, M. Garcia-Closas, R. L. Milne, and J. Chang-Claude. Evidence of gene–environment interactions between common breast cancer susceptibility loci and established environmental risk factors. *PLoS Genet*, 9(3) :e1003284, Mar. 2013.

[93] R. L. Milne, M. M. Gaudet, A. B. Spurdle, P. A. Fasching, F. J. Couch, J. Benítez, J. I. Arias Pérez, M. P. Zamora, N. Malats, I. Dos Santos Silva, L. J. Gibson, O. Fletcher, N. Johnson, H. Anton-Culver, A. Ziogas, J. Figueroa, L. Brinton, M. E. Sherman, J. Lissowska, J. L. Hopper, G. S. Dite, C. Apicella, M. C. Southey, A. J. Sigurdson, M. S. Linet, S. J. Schonfeld, D. M. Freedman, A. Mannermaa, V.-M. Kosma, V. Kataja, P. Auvinen, I. L. Andrulis, G. Glendon, J. A. Knight, N. Weerasooriya, A. Cox, M. W. Reed, S. S. Cross, A. M. Dunning, S. Ahmed, M. Shah, H. Brauch, Y.-D. Ko,

T. Brüning, GENICA Network, D. Lambrechts, J. Reumers, A. Smeets, S. Wang-Gohrke, P. Hall, K. Czene, J. Liu, A. K. Irwanto, G. Chenevix-Trench, H. Holland, kConFab, AOCS, G. G. Giles, L. Baglietto, G. Severi, S. E. Bojensen, B. G. Nordestgaard, H. Flyger, E. M. John, D. W. West, A. S. Whittemore, C. Vachon, J. E. Olson, Z. Fredericksen, M. Kosel, R. Hein, A. Vrieling, D. Flesch-Janys, J. Heinz, M. W. Beckmann, K. Heusinger, A. B. Ekici, L. Haeberle, M. K. Humphreys, J. Morrison, D. F. Easton, P. D. Pharoah, M. García-Closas, E. L. Goode, and J. Chang-Claude. Assessing interactions between the associations of common genetic susceptibility variants, reproductive history and body mass index with breast cancer risk in the breast cancer association consortium : a combined case-control study. *Breast cancer research : BCR*, 12(6) :R110, 2010. PMID : 21194473 PMCID : PMC3046455.

[94] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011) :931–945, Oct. 2004. PMID : 15496913.

[95] A. J. Jeffreys, L. Kauppi, and R. Neumann. Intensely punctate meiotic recombination in the class ii region of the major histocompatibility complex. *Nature genetics*, 29(2) :217–222, Oct. 2001. PMID : 11586303.

[96] F. Baudat, J. Buard, C. Grey, A. Fledel-Alon, C. Ober, M. Przeworski, G. Coop, and B. de Massy. Prdm9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science (New York, N.Y.)*, 327(5967) :836–840, Feb. 2010. PMID : 20044539.

[97] L. Ségurel, E. M. Leffler, and M. Przeworski. The case of the fickle fingers : how the prdm9 zinc finger protein specifies meiotic recombination hotspots in humans. *PLoS biology*, 9(12) :e1001211, Dec. 2011. PMID : 22162947 PMCID : PMC3232208.

[98] T. Miyagawa, M. Kawashima, N. Nishida, J. Ohashi, R. Kimura, A. Fujimoto, M. Shimada, S. Morishita, T. Shigeta, L. Lin, S.-C. Hong, J. Faraco, Y.-K. Shin, J.-H. Jeong, Y. Okazaki, S. Tsuji, M. Honda, Y. Honda, E. Mignot, and K. Tokunaga. Variant between cpt1b and chkb associated with susceptibility to narcolepsy. *Nature Genetics*, 40(11) :1324–1328, Nov. 2008.

[99] S. M. Pulst. Genetic linkage analysis. *Archives of neurology*, 56(6) :667–672, June 1999. PMID : 10369304.

[100] J. M. Hall, M. K. Lee, B. Newman, J. E. Morrow, L. A. Anderson, B. Huey, and M. C. King. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science (New York, N.Y.)*, 250(4988) :1684–1689, Dec. 1990. PMID : 2270482.

[101] Y. Miki, J. Swensen, D. Shattuck-Eidens, P. A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L. M. Bennett, and W. Ding. A strong candidate for the breast and ovarian cancer susceptibility gene brca1. *Science (New York, N.Y.)*, 266(5182) :66–71, Oct. 1994. PMID : 7545954.

[102] R. Wooster, S. L. Neuhausen, J. Mangion, Y. Quirk, D. Ford, N. Collins, K. Nguyen, S. Seal, T. Tran, and D. Averill. Localization of a breast cancer susceptibility gene, brca2, to chromosome 13q12-13. *Science (New York, N.Y.)*, 265(5181) :2088–2090, Sept. 1994. PMID : 8091231.

[103] R. Wooster, G. Bignell, J. Lancaster, S. Swift, S. Seal, J. Mangion, N. Collins, S. Gregory, C. Gumbs, and G. Micklem. Identification of the breast cancer susceptibility gene brca2. *Nature*, 378(6559) :789–792, Dec. 1995. PMID : 8524414.

[104] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1) :289–300, 1995.

[105] J. P. T. Higgins and S. G. Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, 21(11) :1539–1558, June 2002. PMID : 12111919.

[106] Y.-N. Xin, Z.-H. Lin, X.-J. Jiang, S.-H. Zhan, Q.-J. Dong, Q. Wang, and S.-Y. Xuan. Specific hla-dqb1 alleles associated with risk for development of hepatocellular carcinoma : A meta-analysis. *World Journal of Gastroenterology : WJG*, 17(17) :2248–2254, May 2011. PMID : 21633537 PMCID : PMC3092879.

[107] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422) :56–65, Nov. 2012.

[108] F. J. Couch and B. L. Weber. Mutations and polymorphisms in the familial early-onset breast cancer (brca1) gene. breast cancer information core. *Human mutation*, 8(1) :8–18, 1996. PMID : 8807330.

[109] P. Kumar, S. Henikoff, and P. C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature protocols*, 4(7) :1073–1081, 2009. PMID : 19561590.

[110] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4) :248–249, Apr. 2010. PMID : 20354512 PMCID : PMC2855889.

[111] A. Cox, A. M. Dunning, M. Garcia-Closas, S. Balasubramanian, M. W. R. Reed, K. A. Pooley, S. Scollen, C. Baynes, B. A. J. Ponder, S. Chanock, J. Lissowska, L. Brinton, B. Peplonska, M. C. Southey, J. L. Hopper, M. R. E. McCredie, G. G. Giles, O. Fletcher, N. Johnson, I. dos Santos Silva, L. Gibson, S. E. Bojesen, B. G. Nordestgaard, C. K. Axelsson, D. Torres, U. Hamann, C. Justenhoven, H. Brauch, J. Chang-Claude, S. Kropp, A. Risch, S. Wang-Gohrke, P. Schürmann, N. Bogdanova, T. Dörk, R. Fagerholm, K. Aaltonen, C. Blomqvist, H. Nevanlinna, S. Seal, A. Renwick, M. R. Stratton, N. Rahman, S. Sangrajrang, D. Hughes, F. Odefrey, P. Brennan, A. B. Spurdle, G. Chenevix-Trench, Kathleen Cunningham Foundation Consortium for Research into Familial Breast Cancer, J. Beesley, A. Mannermaa, J. Hartikainen, V. Kataja, V.-M. Kosma, F. J. Couch, J. E. Olson, E. L. Goode, A. Broeks, M. K. Schmidt, F. B. L. Hogervorst, L. J. Van't Veer, D. Kang, K.-Y. Yoo, D.-Y. Noh, S.-H. Ahn, S. Wedrén, P. Hall, Y.-L. Low, J. Liu, R. L. Milne, G. Ribas, A. Gonzalez-Neira, J. Benitez, A. J. Sigurdson, D. L. Stredrick, B. H. Alexander, J. P. Struewing, P. D. P. Pharoah, D. F. Easton, and Breast Cancer Association Consortium. A common coding variant in casp8 is associated with breast cancer risk. *Nature genetics*, 39(3) :352–358, Mar. 2007. PMID : 17293864.

[112] The Breast and Prostate Cancer Cohort Consortium. A candidate gene approach to searching for low-penetrance breast and prostate cancer genes. *Nature Reviews Cancer*, 5(12) :977–985, Dec. 2005.

[113] R. J. Klein, C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, and J. Hoh. Complement factor h polymorphism in age-related macular degeneration. *Science (New York, N.Y.)*, 308(5720) :385–389, Apr. 2005. PMID : 15761122 PMCID : PMC1512523.

[114] A. D. Coviello, R. Haring, M. Wellons, D. Vaidya, T. Lehtimäki, S. Keildson, K. L. Lunetta, C. He, M. Fornage, V. Lagou, M. Mangino, N. C. Onland-Moret, B. Chen, J. Eriksson, M. Garcia, Y. M. Liu, A. Koster, K. Lohman, L.-P. Lyytikäinen, A.-K. Petersen, J. Prescott, L. Stolk, L. Vandenput, A. R. Wood, W. V. Zhuang, A. Ruokonen, A.-L. Hartikainen, A. Pouta, S. Bandinelli, R. Biffar, G. Brabant, D. G. Cox, Y. Chen, S. Cummings, L. Ferrucci, M. J. Gunter, S. E. Hankinson, H. Martikainen, A. Hofman, G. Homuth, T. Illig, J.-O. Jansson, A. D. Johnson, D. Karasik, M. Karlsson, J. Kettunen, D. P. Kiel, P. Kraft, J. Liu, O. Ljunggren, M. Lorentzon, M. Maggio, M. R. P. Markus, D. Mellström, I. Miljkovic, D. Mirel, S. Nelson, L. Morin Papunen, P. H. M. Peeters, I. Prokopenko, L. Raffel, M. Reincke, A. P. Reiner, K. Rexrode, F. Rivadeneira, S. M. Schwartz, D. Siscovick, N. Soranzo, D. Stöckl, S. Tworoger, A. G. Uitterlinden, C. H. van Gils, R. S. Vasan, H.-E. Wichmann, G. Zhai, S. Bhasin, M. Bidlingmaier, S. J. Chanock, I. De Vivo, T. B. Harris, D. J. Hunter, M. Kähönen, S. Liu, P. Ouyang, T. D. Spector, Y. T. van der Schouw, J. Viikari, H. Wallaschofski, M. I. McCarthy, T. M. Frayling, A. Murray, S. Franks, M.-R. Järvelin, F. H. de Jong, O. Raitakari, A. Teumer, C. Ohlsson, J. M. Murabito, and J. R. B. Perry. A genome-wide association meta-analysis of circulating sex hormone-binding globulin reveals multiple loci implicated in sex steroid hormone regulation. *PLoS genetics*, 8(7) :e1002805, 2012. PMID : 22829776 PMCID : PMC3400553.

[115] M. R. Nelson, K. Bryc, K. S. King, A. Indap, A. R. Boyko, J. Novembre, L. P. Briley, Y. Maruyama, D. M. Waterworth, G. Waeber, P. Vollenweider, J. R. Oksenberg, S. L. Hauser, H. A. Stirnadel, J. S. Kooner, J. C. Chambers, B. Jones, V. Mooser, C. D. Bustamante, A. D. Roses, D. K. Burns, M. G. Ehm, and E. H. Lai. The population reference sample, popres : a resource for population, disease, and pharmacological genetics research. *American journal of human genetics*, 83(3) :347–358, Sept. 2008. PMID : 18760391 PMCID : PMC2556436.

[116] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens, and C. D. Bustamante. Genes mirror geography within europe. *Nature*, 456(7218) :98–101, Nov. 2008.

[117] K. Hao, E. Chudin, D. Greenawalt, and E. E. Schadt. Magnitude of stratification in human populations and impacts on genome wide association studies. *PLoS ONE*, 5(1) :e8695, Jan. 2010.

[118] D. G. Cox and P. Kraft. Quantification of the power of hardy-weinberg equilibrium testing to detect genotyping error. *Human Heredity*, 61(1) :10–14, 2006. PMID : 16514241.

[119] C. Xu, W. Zhou, Y. Wang, and L. Qiao. Hepatitis b virus-induced hepatocellular carcinoma. *Cancer Letters*, 345(2) :216–222, Apr. 2014.

[120] M. Schiffman and N. Wentzensen. Human papillomavirus infection and the multistage carcinogenesis of cervical cancer. *Cancer Epidemiology Biomarkers & Prevention*, 22(4) :553–560, Jan. 2013. PMID : 23549399.

[121] D. Belpomme, P. Irigaray, L. Hardell, R. Clapp, L. Montagnier, S. Epstein, and A. J. Sasco. The multitude and diversity of environmental carcinogens. *Environmental research*, 105(3) :414–429, Nov. 2007. PMID : 17692309.

[122] M. Satoh and T. Kuroiwa. Organization of multiple nucleoids and dna molecules in mitochondria of a human cell. *Experimental Cell Research*, 196(1) :137–140, Sept. 1991.

[123] H.-C. Lee, S.-H. Li, J.-C. Lin, C.-C. Wu, D.-C. Yeh, and Y.-H. Wei. Somatic mutations in the d-loop and decrease in the copy number of mitochondrial dna in human hepatocellular carcinoma. *Mutation research*, 547(1-2) :71–78, Mar. 2004. PMID : 15013701.

[124] L. Fülöp, G. Szanda, B. Enyedi, P. Várnai, and A. Spät. The effect of opa1 on mitochondrial ca2+ signaling. *PLoS ONE*, 6(9) :e25199, Sept. 2011.

[125] M. Picard, O. S. Shirihai, B. J. Gentil, and Y. Burelle. Mitochondrial morphology transitions and functions : implications for retrograde signaling ? *AJP : Regulatory, Integrative and Comparative Physiology*, 304(6) :R393–R406, Mar. 2013.

[126] K. S. Dimmer and L. Scorrano. (de)constructing mitochondria : What for ? *Physiology*, 21(4) :233–241, Jan. 2006. PMID : 16868312.

[127] Y. G. Yoon, M. D. Koob, and Y. H. Yoo. Re-engineering the mitochondrial genomes in mammalian cells. *Anatomy & Cell Biology*, 43(2) :97, 2010.

[128] M. Sato and K. Sato. Maternal inheritance of mitochondrial dna by diverse mechanisms to eliminate paternal mitochondrial dna. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1833(8) :1979–1984, Aug. 2013.

[129] P. Sutovsky. Ubiquitin-dependent proteolysis in mammalian spermatogenesis, fertilization, and sperm quality control : killing three birds with one stone. *Microscopy research and technique*, 61(1) :88–102, May 2003. PMID : 12672125.

[130] M. Schwartz and J. Vissing. Paternal inheritance of mitochondrial dna. *The New England journal of medicine*, 347(8) :576–580, Aug. 2002. PMID : 12192017.

[131] Y. Kraytsberg, M. Schwartz, T. A. Brown, K. Ebralidse, W. S. Kunz, D. A. Clayton, J. Vissing, and K. Khrapko. Recombination of human mitochondrial dna. *Science*, 304(5673) :981–981, May 2004. PMID : 15143273.

[132] E. D. Ladoukakis and A. Eyre-Walker. Evolutionary genetics : Direct evidence of recombination in human mitochondrial dna. *Heredity*, 93(4) :321–321, Aug. 2004.

[133] O. V. Kidgotko, M. Y. Kustova, V. A. Sokolova, M. G. Bass, and V. B. Vasilyev. Transmission of human mitochondrial dna along the paternal lineage in transmitochondrial mice. *Mitochondrion*, 13(4) :330–336, July 2013.

[134] R. W. Taylor and D. M. Turnbull. Mitochondrial dna mutations in human disease. *Nature Reviews Genetics*, 6(5) :389–402, May 2005.

[135] A. Ramos, C. Santos, L. Mateiu, M. d. M. Gonzalez, L. Alvarez, L. Azevedo, A. Amorim, and M. P. Aluja. Frequency and pattern of heteroplasmy in the complete human mitochondrial genome. *PLoS ONE*, 8(10) :e74636, Oct. 2013.

[136] P. M. Smith and R. N. Lightowlers. Altering the balance between healthy and mutated mitochondrial dna. *Journal of inherited metabolic disease*, 34(2) :309–313, Apr. 2011. PMID : 20506041.

[137] I. J. Holt, A. E. Harding, R. K. Petty, and J. A. Morgan-Hughes. A new mitochondrial disease associated with mitochondrial dna heteroplasmy. *American journal of human genetics*, 46(3) :428–433, Mar. 1990. PMID : 2137962 PMCID : PMC1683641.

[138] A. Rasola and P. Bernardi. Mitochondrial permeability transition in ca2+-dependent apoptosis and necrosis. *Cell Calcium*, 50(3) :222–233, Sept. 2011.

[139] K. A. Webster. Mitochondrial membrane permeabilization and cell death during myocardial infarction : roles of calcium and reactive oxygen species. *Future Cardiology*, 8(6) :863–884, Nov. 2012.

[140] G. W. Dorn. Molecular mechanisms that differentiate apoptosis from programmed necrosis. *Toxicologic Pathology*, 41(2) :227–234, Feb. 2013. PMID : 23222994.

[141] N. B. Larsen, M. Rasmussen, and L. J. Rasmussen. Nuclear and mitochondrial dna repair : similar pathways ? *Mitochondrion*, 5(2) :89–108, Apr. 2005. PMID : 16050976.

[142] M. Yu. Somatic mitochondrial dna mutations in human cancers. In Gregory S. Makowski, editor, *Advances in Clinical Chemistry*, volume Volume 57, pages 99–138. Elsevier, 2012.

[143] S. Song, Z. F. Pursell, W. C. Copeland, M. J. Longley, T. A. Kunkel, and C. K. Mathews. Dna precursor asymmetries in mammalian tissue mitochondria and possible contribution to mutagenesis through reduced replication fidelity. *Proceedings of the National Academy of Sciences of the United States of America*, 102(14) :4990–4995, Apr. 2005. PMID : 15784738 PMCID : PMC555996.

[144] L. L. Cavalli-Sforza and M. W. Feldman. The application of molecular genetic approaches to the study of human evolution. *Nature Genetics*, 33 :266–275, Mar. 2003.

[145] S. L. Mitchell, R. Goodloe, K. Brown-Gentry, S. A. Pendergrass, D. G. Murdock, and D. C. Crawford. Characterization of mitochondrial haplogroups in a large population-based sample from the united states. *Human genetics*, Feb. 2014. PMID : 24488180.

[146] A. Achilli, U. A. Perego, H. Lancioni, A. Olivieri, F. Gandini, B. H. Kashani, V. Battaglia, V. Grugni, N. Angerhofer, M. P. Rogers, R. J. Herrera, S. R. Woodward, D. Labuda, D. G. Smith, J. S. Cybulski, O. Semino, R. S. Malhi, and A. Torroni. Reconciling migration models to the americas with the variation of north american native mitogenomes. *Proceedings of the National Academy of Sciences*, 110(35) :14308–14313, Aug. 2013. PMID : 23940335.

[147] M. van Oven and M. Kayser. Updated comprehensive phylogenetic tree of global human mitochondrial dna variation. *Human Mutation*, 30(2) :E386–E394, 2009.

[148] M. van Oven. Phylotree v15. http ://www.phylotree.org/tree/main.htm.

[149] S. Anderson, A. T. Bankier, B. G. Barrell, M. H. L. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. H. Smith, R. Staden, and I. G. Young. Sequence and organization of the human mitochondrial genome. *Nature*, 290(5806) :457–465, Apr. 1981.

[150] R. M. Andrews, I. Kubacka, P. F. Chinnery, R. N. Lightowlers, D. M. Turnbull, and N. Howell. Reanalysis and revision of the cambridge reference sequence for human mitochondrial dna. *Nature genetics*, 23(2) :147, Oct. 1999. PMID : 10508508.

[151] D. M. Behar, M. van Oven, S. Rosset, M. Metspalu, E.-L. Loogväli, N. M. Silva, T. Kivisild, A. Torroni, and R. Villems. A "copernican" reassessment of the human mitochondrial dna tree from its root. *The American Journal of Human Genetics*, 90(4) :675–684, Apr. 2012.

[152] E. Ruiz-Pesini, M. T. Lott, V. Procaccio, J. C. Poole, M. C. Brandon, D. Mishmar, C. Yi, J. Kreuziger, P. Baldi, and D. C. Wallace. An enhanced mitomap with a global mtdna mutational phylogeny. *Nucleic acids research*, 35(Database issue) :D823–828, Jan. 2007. PMID : 17178747 PMCID : PMC1781213.

[153] Mitomap. http ://www.mitomap.org/MITOMAP.

[154] G. Hudson and P. F. Chinnery. Mitochondrial dna polymerase-gamma and human disease. *Human molecular genetics*, 15 Spec No 2 :R244–252, Oct. 2006. PMID : 16987890.

[155] L. S. Kaguni. Dna polymerase gamma, the mitochondrial replicase. *Annual review of biochemistry*, 73 :293–320, 2004. PMID : 15189144.

[156] G. Van Goethem, B. Dermaut, A. Löfgren, J. J. Martin, and C. Van Broeckhoven. Mutation of polg is associated with progressive external ophthalmoplegia characterized by mtdna deletions. *Nature genetics*, 28(3) :211–212, July 2001. PMID : 11431686.

[157] P. Luoma, A. Melberg, J. O. Rinne, J. A. Kaukonen, N. N. Nupponen, R. M. Chalmers, A. Oldfors, I. Rautakorpi, L. Peltonen, K. Majamaa, H. Somer, and A. Suomalainen. Parkinsonism, premature menopause, and mitochondrial dna polymerase gamma mutations : clinical and molecular genetic study. *Lancet*, 364(9437) :875–882, Sept. 2004. PMID : 15351195.

[158] A. T. Pagnamenta, J.-W. Taanman, C. J. Wilson, N. E. Anderson, R. Marotta, A. J. Duncan, M. Bitner-Glindzicz, R. W. Taylor, A. Laskowski, D. R. Thorburn, and S. Rahman. Dominant inheritance of premature ovarian failure associated with mutant mitochondrial dna polymerase gamma. *Human reproduction (Oxford, England)*, 21(10) :2467–2473, Oct. 2006. PMID : 16595552.

[159] A. T. Rovio, D. R. Marchington, S. Donat, H. C. Schuppe, J. Abel, E. Fritsche, D. J. Elliott, P. Laippala, A. L. Ahola, D. McNay, R. F. Harrison, B. Hughes, T. Barrett, D. M. Bailey, D. Mehmet, A. M. Jequier, T. B. Hargreave, S. H. Kao, J. M. Cummins, D. E. Barton, H. J. Cooke, Y. H. Wei, L. Wichmann, J. Poulton, and H. T. Jacobs. Mutations at the mitochondrial dna polymerase (polg) locus associated with male infertility. *Nature genetics*, 29(3) :261–262, Nov. 2001. PMID : 11687794.

[160] M. Jensen, H. Leffers, J. H. Petersen, A. Nyboe Andersen, N. Jørgensen, E. Carlsen, T. K. Jensen, N. E. Skakkebaek, and E. Rajpert-De Meyts. Frequent polymorphism of the mitochondrial dna polymerase gamma gene (polg) in patients with normal spermiograms and unexplained subfertility. *Human reproduction (Oxford, England)*, 19(1) :65–70, Jan. 2004. PMID : 14688158.

[161] A. Lombes, E. Bonilla, and S. Dimauro. Mitochondrial encephalomyopathies. *Revue neurologique*, 145(10) :671–689, 1989. PMID : 2682927.

[162] J. Schmiedel, S. Jackson, J. Schäfer, and H. Reichmann. Mitochondrial cytopathies. *Journal of neurology*, 250(3) :267–277, Mar. 2003. PMID : 12638015.

[163] K. M. Santa. Treatment options for mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke-like episodes (melas) syndrome. *Pharmacotherapy : The Journal of Human Pharmacology and Drug Therapy*, 30(11) :1179–1196, Nov. 2010.

[164] L. A. Bindoff and B. A. Engelsen. Mitochondrial diseases and epilepsy. *Epilepsia*, 53 Suppl 4 :92–97, Sept. 2012. PMID : 22946726.

[165] A. Fryer, R. Appleton, M. G. Sweeney, L. Rosenbloom, and A. E. Harding. Mitochondrial dna 8993 (narp) mutation presenting with a heterogeneous phenotype including 'cerebral palsy'. *Archives of disease in childhood*, 71(5) :419–422, Nov. 1994. PMID : 7529982 PMCID : PMC1030055.

[166] A. Rojo, Y. Campos, J. M. Sánchez, I. Bonaventura, M. Aguilar, A. García, L. González, M. J. Rey, J. Arenas, M. Olivé, and I. Ferrer. Narp-mils syndrome caused by 8993 t>g mitochondrial dna mutation : a clinical, genetic and neuropathological study. *Acta neuropathologica*, 111(6) :610–616, June 2006. PMID : 16525806.

[167] L. Went. Leber hereditary optic neuropathy (lhon) : a mitochondrial disease with unresolved complexities. *Cytogenetic and Genome Research*, 86(2) :153–156, 1999.

[168] P. Yu-Wai-Man, P. G. Griffiths, and P. F. Chinnery. Mitochondrial optic neuropathies –disease mechanisms and therapeutic strategies. *Progress in Retinal and Eye Research*, 30(2) :81–114, Mar. 2011.

[169] D. LEIGH. Subacute necrotizing encephalomyelopathy in an infant. *Journal of neurology, neurosurgery, and psychiatry*, 14(3) :216–221, Aug. 1951. PMID : 14874135 PMCID : PMC499520.

[170] J. Finsterer. Leigh and leigh-like syndrome in children and adults. *Pediatric Neurology*, 39(4) :223–235, Oct. 2008.

[171] M. Phadke, M. R. Lokeshwar, S. Bhutada, C. Tampi, R. Saxena, S. Kohli, and K. N. Shah. Retracted article : Kearns sayre syndrome—case report with review of literature. *The Indian Journal of Pediatrics*, 79(5) :650–654, May 2012.

[172] D. Hanahan and R. A. Weinberg. Hallmarks of cancer : the next generation. *Cell*, 144(5) :646–674, Mar. 2011. PMID : 21376230.

[173] L. Grzybowska-Szatkowska and B. Slaska. Mitochondrial dna and carcinogenesis (review). *Molecular medicine reports*, 6(5) :923–930, Nov. 2012. PMID : 22895648.

[174] W. Habano, T. Sugai, T. Yoshida, and S. Nakamura. Mitochondrial gene mutation, but not large-scale deletion, is a feature of colorectal carcinomas with mitochondrial microsatellite instability. *International journal of cancer. Journal international du cancer*, 83(5) :625–629, Nov. 1999. PMID : 10521798.

[175] M. S. Fliss, H. Usadel, O. L. Caballero, L. Wu, M. R. Buta, S. M. Eleff, J. Jen, and D. Sidransky. Facile detection of mitochondrial dna mutations in tumors and bodily fluids. *Science (New York, N.Y.)*, 287(5460) :2017–2019, Mar. 2000. PMID : 10720328.

[176] L. J. Burgart, J. Zheng, Q. Shu, J. G. Strickler, and D. Shibata. Somatic mitochondrial mutation in gastric cancer. *The American Journal of Pathology*, 147(4) :1105, Oct. 1995. PMID : 7573355.

[177] P.-H. Yin, C.-C. Wu, J.-C. Lin, C.-W. Chi, Y.-H. Wei, and H.-C. Lee. Somatic mutations of mitochondrial genome in hepatocellular carcinoma. *Mitochondrion*, 10(2) :174–182, Mar. 2010. PMID : 20006738.

[178] M. A. C. Dani, S. U. Dani, S. P. G. Lima, A. Martinez, B. M. Rossi, F. Soares, M. A. Zago, and A. J. G. Simpson. Less deltamtdna4977 than normal in various types of tumors suggests that cancer cells are essentially free of this mutation. *Genetics and molecular research : GMR*, 3(3) :395–409, 2004. PMID : 15614730.

[179] C. Ye, Y.-T. Gao, W. Wen, J. P. Breyer, X. O. Shu, J. R. Smith, W. Zheng, and Q. Cai. Association of mitochondrial dna displacement loop (ca)n dinucleotide repeat polymorphism with breast cancer risk and survival among chinese women. *Cancer Epidemiology Biomarkers & Prevention*, 17(8) :2117–2122, Aug. 2008.

[180] K. Futyma, L. Putowski, M. Cybulski, P. Miotla, T. Rechberger, and A. Semczuk. The prevalence of mtdna4977 deletion in primary human endometrial carcinomas and matched control samples. *Oncology reports*, 20(3) :683–688, Sept. 2008. PMID : 18695924.

[181] C. C. Abnet, K. Huppi, A. Carrera, D. Armistead, K. McKenney, N. Hu, Z.-Z. Tang, P. R. Taylor, and S. M. Dawsey. Control region mutations and the 'common deletion' are frequent in the mitochondrial dna of patients with esophageal squamous cell carcinoma. *BMC cancer*, 4 :30, July 2004. PMID : 15230979 PMCID : PMC459226.

[182] B. H. L. Tan, R. J. E. Skipworth, N. A. Stephens, N. M. Wheelhouse, H. Gilmour, A. C. de Beaux, S. Paterson-Brown, K. C. H. Fearon, and J. A. Ross. Frequency of the mitochondrial dna 4977bp deletion in oesophageal mucosa during the progression of barrett's oesophagus. *European journal of cancer (Oxford, England : 1990)*, 45(5) :736–740, Mar. 2009. PMID : 19211242.

[183] V. Máximo, P. Soares, R. Seruca, A. S. Rocha, P. Castro, and M. Sobrinho-Simões. Microsatellite instability, mitochondrial dna large deletions, and mitochondrial dna mutations in gastric carcinoma. *Genes, chromosomes & cancer*, 32(2) :136–143, Oct. 2001. PMID : 11550281.

[184] P. H. Yin, H. C. Lee, G. Y. Chau, Y. T. Wu, S. H. Li, W. Y. Lui, Y. H. Wei, T. Y. Liu, and C. W. Chi. Alteration of the copy number and deletion of mitochondrial dna in human hepatocellular carcinoma. *British Journal of Cancer*, 90(12) :2390–2396, May 2004.

[185] K. Kotake, T. Nonami, T. Kurokawa, A. Nakao, T. Murakami, and Y. Shimomura. Human livers with cirrhosis and hepatocellular carcinoma have less mitochondrial dna deletion than normal human livers. *Life sciences*, 64(19) :1785–1791, 1999. PMID : 10353633.

[186] N. M. Wheelhouse, P. B. S. Lai, S. J. Wigmore, J. A. Ross, and D. J. Harrison. Mitochondrial d-loop mutations and deletion profiles of cancerous and noncancerous liver tissue in hepatitis b virus-infected liver. *British journal of cancer*, 92(7) :1268–1272, Apr. 2005. PMID : 15785740 PMCID : PMC2361973.

[187] M. Poetsch, A. Petersmann, E. Lignitz, and B. Kleist. Relationship between mitochondrial dna instability, mitochondrial dna large deletions, and nuclear microsatellite instability in head and neck squamous cell carcinomas. *Diagnostic molecular pathology : the American journal of surgical pathology, part B*, 13(1) :26–32, Mar. 2004. PMID : 15163006.

[188] H. C. Lee, P. H. Yin, T. N. Yu, Y. D. Chang, W. C. Hsu, S. Y. Kao, C. W. Chi, T. Y. Liu, and Y. H. Wei. Accumulation of mitochondrial dna deletions in human oral tissues – effects of betel quid chewing and oral cancer. *Mutation research*, 493(1-2) :67–74, June 2001. PMID : 11516716.

[189] J. J. Yu and T. Yan. Effect of mtdna mutation on tumor malignant degree in patients with prostate cancer. *The aging male : the official journal of the International Society for the Study of the Aging Male*, 13(3) :159–165, Sept. 2010. PMID : 20136572.

[190] C. Y. Pang, H. C. Lee, J. H. Yang, and Y. H. Wei. Human skin mitochondrial dna deletions associated with light exposure. *Archives of biochemistry and biophysics*, 312(2) :534–538, Aug. 1994. PMID : 8037468.

[191] G. Tallini, M. Ladanyi, J. Rosai, and S. C. Jhanwar. Analysis of nuclear and mitochondrial dna alterations in thyroid and renal oncocytic tumors. *Cytogenetics and cell genetics*, 66(4) :253–259, 1994. PMID : 7909283.

[192] T. I. Rogounovitch, V. A. Saenko, Y. Shimizu-Yoshida, A. Y. Abrosimov, E. F. Lushnikov, P. O. Roumiantsev, A. Ohtsuru, H. Namba, A. F. Tsyb, and S. Yamashita. Large deletions in mitochondrial dna in radiation-associated human thyroid tumors. *Cancer research*, 62(23) :7031–7041, Dec. 2002. PMID : 12460924.

[193] M. Jastroch, A. S. Divakaruni, S. Mookerjee, J. R. Treberg, and M. D. Brand. Mitochondrial proton and electron leaks. *Essays in Biochemistry*, 47(1) :53–67, June 2010.

[194] B. M. Babior. Nadph oxidase. *Current opinion in immunology*, 16(1) :42–47, Feb. 2004. PMID : 14734109.

[195] E. P. A. Neve and M. Ingelman-Sundberg. Cytochrome p450 proteins : retention and distribution from the endoplasmic reticulum. *Current opinion in drug discovery & development*, 13(1) :78–85, Jan. 2010. PMID : 20047148.

[196] M. Fransen, M. Nordgren, B. Wang, and O. Apanasets. Role of peroxisomes in ros/rns-metabolism : Implications for human disease. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1822(9) :1363–1373, Sept. 2012.

[197] L. Risom, P. Møller, and S. Loft. Oxidative stress-induced dna damage by particulate air pollution. *Mutation research*, 592(1-2) :119–137, Dec. 2005. PMID : 16085126.

[198] A. J. Ghio, M. S. Carraway, and M. C. Madden. Composition of air pollution particles and oxidative stress in cells, tissues, and living systems. *Journal of toxicology and environmental health. Part B, Critical reviews*, 15(1) :1–21, 2012. PMID : 22202227.

[199] J. E. Lee, J. S. Kang, Y.-W. Ki, J. H. Park, I. C. Shin, and H. C. Koh. Fluazinam targets mitochondrial complex i to induce reactive oxygen species-dependent cytotoxicity in sh-sy5y cells. *Neurochemistry International*, 60(8) :773–781, June 2012.

[200] S. J. Stohs, D. Bagchi, E. Hassoun, and M. Bagchi. Oxidative mechanisms in the toxicity of chromium and cadmium ions. *Journal of environmental pathology, toxicology and oncology : official organ of the International Society for Environmental Toxicology and Cancer*, 20(2) :77–88, 2001. PMID : 11394715.

[201] K. H. Al-Gubory. Environmental pollutants and lifestyle factors induce oxidative stress and poor prenatal development. *Reproductive biomedicine online*, Mar. 2014. PMID : 24813750.

[202] T. Nakayama, D. F. Church, and W. A. Pryor. Quantitative analysis of the hydrogen peroxide formed in aqueous cigarette tar extracts. *Free Radical Biology and Medicine*, 7(1) :9–15, 1989.

[203] Y. ISRAEL, M. Rivera-Meza, E. Karahanian, M. E. Quintanilla, L. Tampier, P. Morales, and M. Herrera-Marschitz. Gene specific modifications unravel ethanol and acetaldehyde actions. *Frontiers in Behavioral Neuroscience*, 7 :80, 2013.

[204] I. N. Zelko, T. J. Mariani, and R. J. Folz. Superoxide dismutase multigene family : a comparison of the cuzn-sod (sod1), mn-sod (sod2), and ec-sod (sod3) gene structures, evolution, and expression. *Free radical biology & medicine*, 33(3) :337–349, Aug. 2002. PMID : 12126755.

[205] M. Pinto, J. Neves, M. Palha, and M. Bicho. Oxidative stress in portuguese children with down syndrome. *Down Syndrome Research and Practice*, 8(2) :79–82, 2002.

[206] S. Toppo, L. Flohé, F. Ursini, S. Vanin, and M. Maiorino. Catalytic mechanisms and specificities of glutathione peroxidases : Variations of a basic scheme. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1790(11) :1486–1500, Nov. 2009.

[207] G. C. MILLS. Hemoglobin catabolism. i. glutathione peroxidase, an erythrocyte enzyme which protects hemoglobin from oxidative breakdown. *The Journal of biological chemistry*, 229(1) :189–197, Nov. 1957. PMID : 13491573.

[208] R. Brigelius-Flohé and A. P. Kipp. Physiological functions of gpx2 and its role in inflammation-triggered carcinogenesis. *Annals of the New York Academy of Sciences*, 1259 :19–25, July 2012. PMID : 22758632.

[209] A. Taylor, A. Robson, B. C. Houghton, C. A. Jepson, W. C. L. Ford, and J. Frayne. Epididymal specific, selenium-independent gpx5 protects cells from oxidative stress-induced lipid peroxidation and dna mutation. *Human reproduction (Oxford, England)*, 28(9) :2332–2342, Sept. 2013. PMID : 23696541.

[210] L. Flohé. The fairytale of the gssg/gsh redox potential. *Biochimica et biophysica acta*, 1830(5) :3139–3142, May 2013. PMID : 23127894.

[211] M. Carocho and I. C. F. R. Ferreira. A review on antioxidants, prooxidants and related controversy : Natural and synthetic compounds, screening and analysis methodologies and future perspectives. *Food and Chemical Toxicology*, 51 :15–25, Jan. 2013.

[212] M. L. Wahlqvist. Antioxidant relevance to human health. *Asia Pacific journal of clinical nutrition*, 22(2) :171–176, 2013. PMID : 23635359.

[213] M. L. Hegde, T. Izumi, and S. Mitra. Chapter 6 - oxidized base damage and single-strand break repair in mammalian genomes : Role of disordered regions and posttranslational modifications in early enzymes. In Paul W. Doetsch, editor, *Progress in Molecular Biology and Translational Science*, volume Volume 110 of *Mechanisms of DNA Repair*, pages 123–153. Academic Press, 2012.

[214] J. Cadet, T. Delatour, T. Douki, D. Gasparutto, J. P. Pouget, J. L. Ravanat, and S. Sauvaigo. Hydroxyl radicals and dna base damage. *Mutation research*, 424(1-2) :9–21, Mar. 1999. PMID : 10064846.

[215] M. S. Cooke, M. D. Evans, M. Dizdaroglu, and J. Lunec. Oxidative dna damage : mechanisms, mutation, and disease. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 17(10) :1195–1214, July 2003. PMID : 12832285.

[216] J. A. Petros. mtdna mutations increase tumorigenicity in prostate cancer. *Proceedings of the National Academy of Sciences*, 102(3) :719–724, Jan. 2005.

[217] K. Ishikawa, K. Takenaga, M. Akimoto, N. Koshikawa, A. Yamaguchi, H. Imanishi, K. Nakada, Y. Honma, and J.-I. Hayashi. Ros-generating mitochondrial dna mutations can regulate tumor cell metastasis. *Science*, 320(5876) :661–664, May 2008.

[218] J. M. McCord. The evolution of free radicals and oxidative stress. *The American journal of medicine*, 108(8) :652–659, June 2000. PMID : 10856414.

[219] R. S. Arnold, C. Q. Sun, J. C. Richards, G. Grigoriev, I. M. Coleman, P. S. Nelson, C.-L. Hsieh, J. K. Lee, Z. Xu, A. Rogatko, A. O. Osunkoya, M. Zayzafoon, L. Chung, and J. A. Petros. Mitochondrial dna mutation stimulates prostate cancer growth in bone stromal environment. *The Prostate*, 69(1) :1–11, Jan. 2009.

[220] M. Mattiazzi, C. Vijayvergiya, C. D. Gajewski, D. C. DeVivo, G. Lenaz, M. Wiedmann, and G. Manfredi. The mtdna t8993g (narp) mutation results in an impairment of oxidative phosphorylation that can be improved by antioxidants. *Human molecular genetics*, 13(8) :869–879, Apr. 2004. PMID : 14998933.

[221] J. A. Canter, A. R. Kallianpur, F. F. Parl, and R. C. Millikan. Mitochondrial dna g10398a polymorphism and invasive breast cancer in african-american women. *Cancer research*, 65(17) :8028–8033, Sept. 2005. PMID : 16140977.

[222] K. Darvishi, S. Sharma, A. K. Bhat, E. Rai, and R. N. K. Bamezai. Mitochondrial dna g10398a polymorphism imparts maternal haplogroup n a risk for breast and esophageal cancer. *Cancer letters*, 249(2) :249–255, May 2007. PMID : 17081685.

[223] V. W. Setiawan, L.-H. Chu, E. M. John, Y. C. Ding, S. A. Ingles, L. Bernstein, M. F. Press, G. Ursin, C. A. Haiman, and S. L. Neuhausen. Mitochondrial dna g10398a variant is not associated with breast cancer in african-american women. *Cancer genetics and cytogenetics*, 181(1) :16–19, Feb. 2008. PMID : 18262047 PMCID : PMC3225405.

[224] A. Francis, S. Pooja, S. Rajender, P. Govindaraj, N. R. Tipirisetti, D. Surekha, D. R. Rao, L. Rao, L. Ramachandra, S. Vishnupriya, K. Ramalingam, K. Satyamoorthy, and K. Thangaraj. A mitochondrial dna variant 10398g>a in breast cancer among south indians : an original study with meta-analysis. *Mitochondrion*, 13(6) :559–565, Nov. 2013. PMID : 23993954.

[225] R.-K. Bai, S. M. Leal, D. Covarrubias, A. Liu, and L.-J. C. Wong. Mitochondrial genetic background modifies breast cancer risk. *Cancer research*, 67(10) :4687–4694, May 2007. PMID : 17510395.

[226] A. M. Czarnecka, J. S. Czarnecki, W. Kukwa, F. Cappello, A. Ścińska, and A. Kukwa. Molecular oncology focus - is carcinogenesis a 'mitochondriopathy' ? *Journal of Biomedical Science*, 17(1) :31, Apr. 2010. PMID : 20416110.

[227] N. Tengku Baharudin, H. Jaafar, and Z. Zainuddin. Association of mitochondrial dna 10398 polymorphism in invasive breast cancer in malay population of peninsular malaysia.

The Malaysian journal of medical sciences : MJMS, 19(1) :36–42, Jan. 2012. PMID : 22977373 PMCID : PMC3436496.

[228] D. Covarrubias, R.-K. Bai, L.-J. C. Wong, and S. M. Leal. Mitochondrial dna variant interactions modify breast cancer risk. *J Hum Genet*, 53(10) :924–928, 2008.

[229] A. Pezzotti, P. Kraft, S. E. Hankinson, D. J. Hunter, J. Buring, and D. G. Cox. The mitochondrial a10398g polymorphism, interaction with alcohol consumption, and breast cancer risk. *PLoS One*, 4(4) :e5356, 2009.

[230] A. Bhat, A. Koul, S. Sharma, E. Rai, S. I. A. Bukhari, M. K. Dhar, and R. N. K. Bamezai. The possible role of 10398a and 16189c mtdna variants in providing susceptibility to t2dm in two north indian populations : a replicative study. *Human genetics*, 120(6) :821–826, Feb. 2007. PMID : 17066297.

[231] Y. Wang, V. W. S. Liu, P. C. K. Tsang, P. M. Chiu, A. N. Y. Cheung, U. S. Khoo, P. Nagley, and H. Y. S. Ngan. Microsatellite instability in mitochondrial genome of common female cancers. *International journal of gynecological cancer : official journal of the International Gynecological Cancer Society*, 16 Suppl 1 :259–266, Feb. 2006. PMID : 16515601.

[232] A. M. Czarnecka, T. Krawczyk, K. Plak, A. Klemba, M. Zdrozny, R. S. Arnold, B. Kofler, P. Golik, A. Szybinska, J. Lubinski, M. Mossakowska, E. Bartnik, and J. A. Petros. Mitochondrial genotype and breast cancer predisposition. *Oncol Rep*, 24(6) :1521–1534, Dec. 2010.

[233] N. R. Tipirisetti, R. K. Lakshmi, S. Govatati, S. Govatati, S. Vuree, L. Singh, D. Raghunadha Rao, M. Bhanoori, and S. Vishnupriya. Mitochondrial genome variations in advanced stage breast cancer : a case-control study. *Mitochondrion*, 13(4) :372–378, July 2013. PMID : 23628690.

[234] Y. Chen and J. Pei. Possible risk modifications in the association between mnsod ala-9val polymorphism and breast cancer risk : subgroup analysis and evidence-based sample size calculation for a future trial. *Breast Cancer Res Treat*, 125(2) :495–504, Jan. 2011.

[235] S. Ezzikouri, A. E. El Feydi, R. Afifi, M. Benazzouz, M. Hassar, P. Pineau, and S. Benjelloun. Polymorphisms in antioxidant defence genes and susceptibility to hepatocellular carcinoma in a moroccan population. *Free Radic Res*, 44(2) :208–216, Feb. 2010.

[236] T. Pietras, J. Szemraj, A. Witusik, M. Ho ?ub, M. Panek, R. Wujcik, and P. G—rski. The sequence polymorphism of mnsod gene in subjects with respiratory insufficiency in copd. *Med Sci Monit*, 16(9) :CR427–CR432, Sept. 2010.

[237] D. G. Cox, R. M. Tamimi, and D. J. Hunter. Gene x gene interaction between mnsod and gpx-1 and breast cancer risk : a nested case-control study. *BMC Cancer*, 6 :217, 2006.

[238] D. J. Hunter, E. Riboli, C. A. Haiman, D. Albanes, D. Altshuler, S. J. Chanock, R. B. Haynes, B. E. Henderson, R. Kaaks, D. O. Stram, G. Thomas, M. J. Thun, H. Blanché, J. E. Buring, N. P. Burtt, E. E. Calle, H. Cann, F. Canzian, Y. C. Chen, G. A. Colditz,

D. G. Cox, A. M. Dunning, H. S. Feigelson, M. L. Freedman, J. M. Gaziano, E. Giovannucci, S. E. Hankinson, J. N. Hirschhorn, R. N. Hoover, T. Key, L. N. Kolonel, P. Kraft, L. Le Marchand, S. Liu, J. Ma, S. Melnick, P. Pharaoh, M. C. Pike, C. Rodriguez, V. W. Setiawan, M. J. Stampfer, E. Trapido, R. Travis, J. Virtamo, S. Wacholder, W. C. Willett, and National Cancer Institute Breast and Prostate Cancer Cohort Consortium. A candidate gene approach to searching for low-penetrance breast and prostate cancer genes. *Nature reviews. Cancer*, 5(12) :977–985, Dec. 2005. PMID : 16341085.

[239] S. J. Hendrickson, S. Lindström, A. H. Eliassen, B. A. Rosner, C. Chen, M. Barrdahl, L. Brinton, J. Buring, F. Canzian, S. Chanock, F. Clavel-Chapelon, J. D. Figueroa, S. M. Gapstur, M. Garcia-Closas, M. M. Gaudet, C. A. Haiman, A. Hazra, B. Henderson, R. Hoover, A. Hüsing, M. Johansson, R. Kaaks, K.-T. Khaw, L. N. Kolonel, L. Le Marchand, J. Lissowska, E. Lund, M. L. McCullough, B. Peplonska, E. Riboli, C. Sacerdote, M.-J. Sánchez, A. Tjønneland, D. Trichopoulos, C. H. van Gils, M. Yeager, P. Kraft, D. J. Hunter, R. G. Ziegler, and W. C. Willett. Plasma carotenoid- and retinol-weighted multi-snp scores and risk of breast cancer in the national cancer institute breast and prostate cancer cohort consortium. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 22(5) :927–936, May 2013. PMID : 23515144 PMCID : PMC3650115.

[240] W. J. Gauderman. Sample size requirements for matched case-control studies of gene-environment interaction. *Stat Med*, 21(1) :35–50, Jan. 2002.

[241] O. Borgan. Modeling survival data : Extending the cox model. terry m. therneau and patricia m. grambsch, springer-verlag, new york, 2000. no. of pages : xiii + 350. price : $69.95. isbn 0-387-98784-3. *Statistics in Medicine*, 20(13) :2053–2054, 2001.

[242] Z. Arsova-Sarafinovska, N. Matevska, A. Eken, D. Petrovski, S. Banev, S. Dzikova, V. Georgiev, A. Sikole, O. Erdem, A. Sayal, A. Aydin, and A. J. Dimovski. Glutathione peroxidase 1 (gpx1) genetic polymorphism, erythrocyte gpx activity, and prostate cancer risk. *International urology and nephrology*, 41(1) :63–70, 2009. PMID : 18563616.

[243] T.-Y. D. Cheng, M. J. Barnett, A. R. Kristal, C. B. Ambrosone, I. B. King, M. D. Thornquist, G. E. Goodman, and M. L. Neuhouser. Genetic variation in myeloperoxidase modifies the association of serum α-tocopherol with aggressive prostate cancer among current smokers. *The Journal of nutrition*, 141(9) :1731–1737, Sept. 2011. PMID : 21795425.

[244] J.-Y. Choi, M. L. Neuhouser, M. Barnett, M. Hudson, A. R. Kristal, M. Thornquist, I. B. King, G. E. Goodman, and C. B. Ambrosone. Polymorphisms in oxidative stress–related genes are not associated with prostate cancer risk in heavy smokers. *Cancer Epidemiology Biomarkers & Prevention*, 16(6) :1115–1120, Jan. 2007.

[245] O. Erdem, A. Eken, C. Akay, Z. Arsova-Sarafinovska, N. Matevska, L. Suturkova, K. Erten, Y. Özgök, A. Dimovski, A. Sayal, and A. Aydin. Association of gpx1 polymorphism, gpx activity and prostate cancer risk. *Human & experimental toxicology*, 31(1) :24–31, Jan. 2012. PMID : 21636625.

[246] C. Kucukgergin, M. Gokpinar, O. Sanli, T. Tefik, T. Oktar, and S. Seckin. Association between genetic variants in glutathione peroxidase 1 (gpx1) gene, gpx activity and the risk of prostate cancer. *Minerva urologica e nefrologica = The Italian journal of urology and nephrology*, 63(3) :183–190, Sept. 2011. PMID : 21993316.

[247] A. Steinbrecher, C. Méplan, J. Hesketh, L. Schomburg, T. Endermann, E. Jansen, B. Akesson, S. Rohrmann, and J. Linseisen. Effects of selenium status and polymorphisms in selenoprotein genes on prostate cancer risk in a prospective study of european men. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 19(11) :2958–2968, Nov. 2010. PMID : 20852007.

[248] T. Men, X. Zhang, J. Yang, B. Shen, X. Li, D. Chen, and J. Wang. The rs1050450 c > t polymorphism of gpx1 is associated with the risk of bladder but not prostate cancer : evidence from a meta-analysis. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine*, 35(1) :269–275, Jan. 2014. PMID : 23975365.

[249] D. J. Baer, J. T. Judd, B. A. Clevidence, R. A. Muesing, W. S. Campbell, E. D. Brown, and P. R. Taylor. Moderate alcohol consumption lowers risk factors for cardiovascular disease in postmenopausal women fed a controlled diet. *The American journal of clinical nutrition*, 75(3) :593–599, Mar. 2002. PMID : 11864868.

[250] S. S. Hellmann, L. C. Thygesen, J. S. Tolstrup, and M. Grønbaek. Modifiable risk factors and survival in women diagnosed with primary breast cancer : results from a prospective cohort study. *European journal of cancer prevention : the official journal of the European Cancer Prevention Organisation (ECP)*, 19(5) :366–373, Sept. 2010. PMID : 20502344.

[251] C. Stefanadis, A. Synetos, D. Tousoulis, E. Tsiamis, A. Michelongona, F. Zagouri, A. Bamias, M. A. Dimopoulos, S. Kyvelou, I. Kapelakis, and K. Toutouzas. Systemic administration of bevacizumab increases the risk of cardiovascular events in patients with metastatic cancer. *International journal of cardiology*, 154(3) :341–344, Feb. 2012. PMID : 22078988.

[252] L. W. Jones, M. Haykowsky, C. J. Peddle, A. A. Joy, E. N. Pituskin, L. M. Tkachuk, K. S. Courneya, D. J. Slamon, and J. R. Mackey. Cardiovascular risk profile of patients with her2/neu-positive breast cancer treated with anthracycline-taxane-containing adjuvant chemotherapy and/or trastuzumab. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 16(5) :1026–1031, May 2007. PMID : 17507633.

[253] M. Bonifazi, M. Franchi, M. Rossi, L. Moja, A. Zambelli, A. Zambon, G. Corrao, C. La Vecchia, C. Zocchetti, and E. Negri. Trastuzumab-related cardiotoxicity in early breast cancer : a cohort study. *The oncologist*, 18(7) :795–801, 2013. PMID : 23823908 PMCID : PMC3720632.

[254] L. C. Magnano, N. Martínez Cibrian, X. Andrade González, and X. Bosch. Cardiac complications of chemotherapy : role of prevention. *Current treatment options in cardiovascular medicine*, 16(6) :312, June 2014. PMID : 24817319.

[255] J. Lemieux, C. Diorio, M.-A. Côté, L. Provencher, F. Barabé, S. Jacob, C. St-Pierre, E. Demers, R. Tremblay-Lemay, C. Nadeau-Larochelle, A. Michaud, and C. Laflamme. Alcohol and her2 polymorphisms as risk factor for cardiotoxicity in breast cancer treated with trastuzumab. *Anticancer research*, 33(6) :2569–2576, June 2013. PMID : 23749910.

[256] L. Roca, V. Diéras, H. Roché, E. Lappartient, P. Kerbrat, L. Cany, S. Chieze, J.-L. Canon, M. Spielmann, F. Penault-Llorca, A.-L. Martin, C. Mesleard, J. Lemonnier, and P. de Cremoux. Correlation of her2, fcgr2a, and fcgr3a gene polymorphisms with trastuzumab related cardiac toxicity and efficacy in a subgroup of patients from unicancer-pacs 04 trial. *Breast cancer research and treatment*, 139(3) :789–800, June 2013. PMID : 23780683.

[257] F. S. M. Hilbers, N. B. Boekel, A. J. van den Broek, R. van Hien, S. Cornelissen, B. M. P. Aleman, L. J. van 't Veer, F. E. van Leeuwen, and M. K. Schmidt. Genetic variants in tgfβ-1 and pai-1 as possible risk factors for cardiovascular disease after radiotherapy for breast cancer. *Radiotherapy and Oncology : Journal of the European Society for Therapeutic Radiology and Oncology*, 102(1) :115–121, Jan. 2012. PMID : 22100658.

[258] S. Blein, S. Berndt, A. D. Joshi, D. Campa, R. G. Ziegler, E. Riboli, D. G. Cox, and on Behalf of the NCI Breast and Prostate Cancer Cohort Consortium. Factors associated with oxidative stress and cancer risk in the breast and prostate cancer cohort consortium. *Free radical research*, Jan. 2014. PMID : 24437375.

[259] O. Bahcall. Cogs project and design of the icogs array. *Nature Genetics*, 2013.

[260] Breast Cancer Association Consortium. Commonly studied single-nucleotide polymorphisms and breast cancer : results from the breast cancer association consortium. *Journal of the National Cancer Institute*, 98(19) :1382–1396, Oct. 2006. PMID : 17018785.

[261] The breast cancer association consortium. http ://ccge.medschl.cam.ac.uk/consortia/bcac/, 2005.

[262] A. Berchuck, J. M. Schildkraut, C. L. Pearce, G. Chenevix-Trench, and P. D. Pharoah. Role of genetic polymorphisms in ovarian cancer susceptibility : Development of an international ovarian cancer association consortium. In G. Coukos, A. Berchuck, and R. Ozols, editors, *Ovarian Cancer*, number 622 in Advances in Experimental Medicine and Biology, pages 53–67. Springer New York, Jan. 2008.

[263] The ovarian cancer association consortium. http ://ccge.medschl.cam.ac.uk/consortia/ocac/.

[264] Z. Kote-Jarai, D. F. Easton, J. L. Stanford, E. A. Ostrander, J. Schleutker, S. A. Ingles, D. Schaid, S. Thibodeau, T. Dörk, D. Neal, J. Donovan, F. Hamdy, A. Cox, C. Maier, W. Vogel, M. Guy, K. Muir, A. Lophatananon, M.-A. Kedda, A. Spurdle, S. Steginga, E. M. John, G. Giles, J. Hopper, P. O. Chappuis, P. Hutter, W. D. Foulkes, N. Hamel, C. A. Salinas, J. S. Koopmeiners, D. M. Karyadi, B. Johanneson, T. Wahlfors, T. L. Tammela, M. C. Stern, R. Corral, S. K. McDonnell, P. Schürmann, A. Meyer, R. Kuefer, D. A. Leongamornlert, M. Tymrakiewicz, J.-F. Liu, T. O'Mara, R. A. F. Gardiner, J. Aitken, A. D. Joshi, G. Severi, D. R. English, M. Southey, S. M. Edwards, A. A. Al Olama,

PRACTICAL Consortium, and R. A. Eeles. Multiple novel prostate cancer predisposition loci confirmed by an international study : the practical consortium. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 17(8) :2052–2061, Aug. 2008. PMID : 18708398 PMCID : PMC2776652.

[265] Prostate cancer association group to investigate cancer associated alterations in the genome. http ://ccge.medschl.cam.ac.uk/consortia/practical/.

[266] G. Chenevix-Trench, R. L. Milne, A. C. Antoniou, F. J. Couch, D. F. Easton, D. E. Goldgar, and CIMBA. An international initiative to identify genetic modifiers of cancer risk in brca1 and brca2 mutation carriers : the consortium of investigators of modifiers of brca1 and brca2 (cimba). *Breast cancer research : BCR*, 9(2) :104, 2007. PMID : 17466083 PMCID : PMC1868919.

[267] The consortium of investigators of modifiers of brca1/2. http ://ccge.medschl.cam.ac.uk/consortia/cimba/index.html.

[268] V. Alvarez-Iglesias, A. Mosquera-Miguel, M. Cerezo, B. Quintáns, M. T. Zarrabeitia, I. Cuscó, M. V. Lareu, O. García, L. Pérez-Jurado, A. Carracedo, and A. Salas. New population and phylogenetic features of the internal variation within mitochondrial dna macrohaplogroup r0. *PloS one*, 4(4) :e5112, 2009. PMID : 19340307 PMCID : PMC2660437.

[269] M. Ingman and U. Gyllensten. mtdb : Human mitochondrial genome database, a resource for population genetics and medical sciences. *Nucleic Acids Research*, 34(suppl 1) :D749–D751, Jan. 2006. PMID : 16381973.

[270] C. Bardel, V. Danjean, and E. Génin. Altree : association detection and localization of susceptibility sites using haplotype phylogenetic trees. *Bioinformatics*, 22(11) :1402–1403, June 2006.

[271] C. Bardel, V. Danjean, P. Morange, E. Génin, and P. Darlu. On the use of phylogeny-based tests to detect association between quantitative traits and haplotypes. *Genetic epidemiology*, 33(8) :729–739, Dec. 2009. PMID : 19399905.

[272] C. Bardel. Utilisation de phylogénies d'haplotypes pour la mise en évidence de facteurs génétiques de risque dans les maladies complexe. Master's thesis, Université Paris VI, July 2002.

[273] Y. Ge, S. Dudoit, and T. P. Speed. Resampling-based multiple testing for microarray data analysis. *Test*, 12(1) :1–77, June 2003.

[274] Z. Yang. Paml 4 : phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8) :1586–1591, Aug. 2007. PMID : 17483113.

[275] A. C. Antoniou, P. D. P. Pharoah, S. Narod, H. A. Risch, J. E. Eyfjord, J. L. Hopper, H. Olsson, O. Johannsson, A. Borg, B. Pasini, P. Radice, S. Manoukian, D. M. Eccles, N. Tang, E. Olah, H. Anton-Culver, E. Warner, J. Lubinski, J. Gronwald, B. Gorski, H. Tulinius, S. Thorlacius, H. Eerola, H. Nevanlinna, K. Syrjäkoski, O.-P. Kallioniemi,

D. Thompson, C. Evans, J. Peto, F. Lalloo, D. G. Evans, and D. F. Easton. Breast and ovarian cancer risks to carriers of the brca1 5382insc and 185delag and brca2 6174delt mutations : a combined analysis of 22 population based studies. *Journal of Medical Genetics*, 42(7) :602–603, Jan. 2005. PMID : 15994883.

[276] A. Y. Shuen and W. D. Foulkes. Inherited mutations in breast cancer genes—risk and response. *Journal of Mammary Gland Biology and Neoplasia*, 16(1) :3–15, Apr. 2011.

[277] H.-J. Bandelt, A. Kloss-Brandstätter, M. B. Richards, Y.-G. Yao, and I. Logan. The case for the continuing use of the revised cambridge reference sequence (rcrs) and the standardization of notation in human mitochondrial dna studies. *Journal of Human Genetics*, Dec. 2013.

[278] G. Jönsson, T. L. Naylor, J. Vallon-Christersson, J. Staaf, J. Huang, M. R. Ward, J. D. Greshock, L. Luts, H. Olsson, N. Rahman, M. Stratton, M. Ringnér, A. Borg, and B. L. Weber. Distinct genomic profiles in hereditary breast tumors identified by array-based comparative genomic hybridization. *Cancer research*, 65(17) :7612–7621, Sept. 2005. PMID : 16140926.

[279] K. W. Caldecott. Single-strand break repair and genetic disease. *Nature Reviews Genetics*, 9(8) :619–631, Aug. 2008.

[280] A. A. Davies, J.-Y. Masson, M. J. McIlwraith, A. Z. Stasiak, A. Stasiak, A. R. Venkitaraman, and S. C. West. Role of brca2 in control of the rad51 recombination and dna repair protein. *Molecular Cell*, 7(2) :273–282, Feb. 2001.

[281] W.-T. Zhao, Y.-T. Wang, Z.-W. Huang, and J. Fang. Brca2 affects the efficiency of dna double-strand break repair in response to n-nitroso compounds with differing carcinogenic potentials. *Oncology Letters*, 5(6) :1948–1954, June 2013. PMID : 23833673 PMCID : PMC3700919.

[282] L. Davis and N. Maizels. Homology-directed repair of dna nicks via pathways distinct from canonical double-strand break repair. *Proceedings of the National Academy of Sciences of the United States of America*, 111(10) :E924–932, Mar. 2014. PMID : 24556991 PMCID : PMC3956201.

[283] V. Abkevich, K. M. Timms, B. T. Hennessy, J. Potter, M. S. Carey, L. A. Meyer, K. Smith-McCune, R. Broaddus, K. H. Lu, J. Chen, T. V. Tran, D. Williams, D. Iliev, S. Jammulapati, L. M. FitzGerald, T. Krivak, J. A. DeLoia, A. Gutin, G. B. Mills, and J. S. Lanchbury. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *British journal of cancer*, 107(10) :1776–1782, Nov. 2012. PMID : 23047548 PMCID : PMC3493866.

[284] E. E. Mueller, E. Schaier, S. M. Brunner, W. Eder, J. A. Mayr, S. F. Egger, C. Nischler, H. Oberkofler, H. A. Reitsamer, W. Patsch, W. Sperl, and B. Kofler. Mitochondrial haplogroups and control region polymorphisms in age-related macular degeneration : A case-control study. *PLoS ONE*, 7(2) :e30874, Feb. 2012.

[285] T. Amo, N. Yadava, R. Oh, D. G. Nicholls, and M. D. Brand. Experimental assessment of bioenergetic differences caused by the common european mitochondrial dna haplogroups h and t. *Gene*, 411(1–2) :69–76, Mar. 2008.

[286] P. Parrella, Y. Xiao, M. Fliss, M. Sanchez-Cespedes, P. Mazzarelli, M. Rinaldi, T. Nicol, E. Gabrielson, C. Cuomo, D. Cohen, S. Pandit, M. Spencer, C. Rabitti, V. M. Fazio, and D. Sidransky. Detection of mitochondrial dna mutations in primary breast cancer and fine-needle aspirates. *Cancer Research*, 61(20) :7623–7626, Oct. 2001. PMID : 11606403.

[287] M. Brandon, P. Baldi, and D. C. Wallace. Mitochondrial mutations in cancer. *Oncogene*, 25(34) :4647–4662, Aug. 2006. PMID : 16892079.

[288] F. S. M. Hilbers, M. P. G. Vreeswijk, C. J. van Asperen, and P. Devilee. The impact of next generation sequencing on the analysis of breast cancer susceptibility : a role for extremely rare genetic variation ? *Clinical Genetics*, 84(5) :407–414, Nov. 2013. PMID : 24025038.

[289] COMPLEXO, M. C. Southey, D. J. Park, T. Nguyen-Dumont, I. Campbell, E. Thompson, A. H. Trainer, G. Chenevix-Trench, J. Simard, M. Dumont, P. Soucy, M. Thomassen, L. Jønson, I. S. Pedersen, T. V. Hansen, H. Nevanlinna, S. Khan, O. Sinilnikova, S. Mazoyer, F. Lesueur, F. Damiola, R. Schmutzler, A. Meindl, E. Hahnen, M. R. Dufault, T. Chris Chan, A. Kwong, R. Barkardóttir, P. Radice, P. Peterlongo, P. Devilee, F. Hilbers, J. Benitez, A. Kvist, T. Törngren, D. Easton, D. Hunter, S. Lindstrom, P. Kraft, W. Zheng, Y.-T. Gao, J. Long, S. Ramus, B.-J. Feng, J. N. Weitzel, K. Nathanson, K. Offit, V. Joseph, M. Robson, K. Schrader, S. M. Wang, Y. C. Kim, H. Lynch, C. Snyder, S. Tavtigian, S. Neuhausen, F. J. Couch, and D. E. Goldgar. Complexo : identifying the missing heritability of breast cancer via next generation collaboration. *Breast cancer research : BCR*, 15(3) :402, June 2013. PMID : 23809231 PMCID : PMC3706918.

[290] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England)*, 25(14) :1754–1760, July 2009. PMID : 19451168 PMCID : PMC2705234.

[291] Fastqc project. http ://www.bioinformatics.bbsrc.ac.uk/projects/fastqc.

[292] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics (Oxford, England)*, 25(16) :2078–2079, Aug. 2009. PMID : 19505943 PMCID : PMC2723002.

[293] Picard project. http ://picard.sourceforge.net.

[294] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The genome analysis toolkit : a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9) :1297–1303, Sept. 2010. PMID : 20644199 PMCID : PMC2928508.

[295] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M.

Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5) :491–498, May 2011. PMID : 21478889 PMCID : PMC3083463.

[296] P. Flicek, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. Hunt, N. Johnson, T. Juettemann, A. K. Kähäri, S. Keenan, E. Kulesha, F. J. Martin, T. Maurel, W. M. McLaren, D. N. Murphy, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, M. Ruffier, D. Sheppard, K. Taylor, A. Thormann, S. J. Trevanion, A. Vullo, S. P. Wilder, M. Wilson, A. Zadissa, B. L. Aken, E. Birney, F. Cunningham, J. Harrow, J. Herrero, T. J. P. Hubbard, R. Kinsella, M. Muffato, A. Parker, G. Spudich, A. Yates, D. R. Zerbino, and S. M. J. Searle. Ensembl 2014. *Nucleic Acids Research*, 42(D1) :D749–D755, Jan. 2014. PMID : 24316576.

[297] Ion torrent sequencing on humans.

[298] E. Picardi and G. Pesole. Mitochondrial genomes gleaned from human whole-exome sequencing. *Nature Methods*, 9(6) :523–524, June 2012.

[299] M. Cheng, Z. Guo, H. Li, Z. Li, C. Li, and C. Geng. Identification of sequence polymorphisms in the mitochondrial displacement loop as risk factors for sporadic and familial breast cancer. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine*, Jan. 2014. PMID : 24430364.

[300] H. K. Soini, J. S. Moilanen, T. Vilmi-Kerälä, S. Finnilä, and K. Majamaa. Mitochondrial dna variant m.15218a &gt ; g in finnish epilepsy patients who have maternal relatives with epilepsy, sensorineural hearing impairment or diabetes mellitus. *BMC Medical Genetics*, 14(1) :73, 2013.

[301] F. Legros, E. Chatzoglou, P. Frachon, H. Ogier De Baulny, P. Laforêt, C. Jardel, C. Godinot, and A. Lombès. Functional characterization of novel mutations in the human cytochrome b gene. *European journal of human genetics : EJHG*, 9(7) :510–518, July 2001. PMID : 11464242.

[302] L. M. Bragg, G. Stone, M. K. Butler, P. Hugenholtz, and G. W. Tyson. Shining a light on dark sequencing : Characterising errors in ion torrent pgm data. *PLoS Comput Biol*, 9(4) :e1003031, Apr. 2013.

[303] M. Quail, M. E. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu. A tale of three next generation sequencing platforms : comparison of ion torrent, pacific biosciences and illumina miseq sequencers. *BMC Genomics*, 13(1) :341, 2012.

[304] A. M. Elliott, J. Radecki, B. Moghis, X. Li, and A. Kammesheidt. Rapid detection of the acmg/acog-recommended 23 cftr disease-causing mutations using ion torrent semiconductor sequencing. *Journal of biomolecular techniques : JBT*, 23(1) :24–30, Apr. 2012. PMID : 22468138 PMCID : PMC3313698.

[305] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1) :24–26, Jan. 2011.

[306] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov. Integrative genomics viewer (igv) : high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2) :178–192, Jan. 2013. PMID : 22517427.

[307] N. J. Loman, R. V. Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain, and M. J. Pallen. Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, 30(5) :434–439, May 2012.

[308] W. Parson, C. Strobl, G. Huber, B. Zimmermann, S. M. Gomes, L. Souto, L. Fendt, R. Delport, R. Langit, S. Wootton, R. Lagacé, and J. Irwin. Reprint of : Evaluation of next generation mtgenome sequencing using the ion torrent personal genome machine (pgm). *Forensic Science International : Genetics*, 7(6) :632–639, Dec. 2013.

[309] D. R. Schrider and A. D. Kern. Discovering functional dna elements using population genomic information : A proof of concept using human mtdna. *Genome Biology and Evolution*, 6(7) :1542–1548, Jan. 2014. PMID : 24916662.

[310] M. Goujon, H. McWilliam, W. Li, F. Valentin, S. Squizzato, J. Paern, and R. Lopez. A new bioinformatics analysis tools framework at embl-ebi. *Nucleic Acids Research*, 38(Web Server issue) :W695–699, July 2010. PMID : 20439314 PMCID : PMC2896090.

[311] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology*, 7 :539, 2011. PMID : 21988835 PMCID : PMC3261699.

[312] J. Castresana. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, 17(4) :540–552, Apr. 2000. PMID : 10742046.

[313] U. M. Marigorta and G. Gibson. A simulation study of gene-by-environment interactions in gwas implies ample hidden effects. *Evolutionary and Population Genetics*, 5 :225, 2014.

[314] H. Lynch and H. Wen. Can unknown predisposition in familial breast cancer be family-specific ? *The breast journal*, 2013.

[315] J. W. Yarham, M. Al-Dosary, E. L. Blakely, C. L. Alston, R. W. Taylor, J. L. Elson, and R. McFarland. A comparative analysis approach to determining the pathogenicity of mitochondrial trna mutations. *Human Mutation*, 32(11) :1319–1325, Nov. 2011. PMID : 21882289.

[316] E. González-Vioque, B. Bornstein, M. E. Gallardo, M. A. Fernández-Moreno, and R. Garesse. The pathogenicity scoring system for mitochondrial trna mutations revisited. *Molecular Genetics & Genomic Medicine*, 2(2) :107–114, Mar. 2014.

[317] F. J. Gracia-Aznarez, V. Fernandez, G. Pita, P. Peterlongo, O. Dominguez, M. de la Hoya, M. Duran, A. Osorio, L. Moreno, A. Gonzalez-Neira, J. M. Rosa-Rosa, O. Sinilnikova, S. Mazoyer, J. Hopper, C. Lazaro, M. Southey, F. Odefrey, S. Manoukian, I. Catucci, T. Caldes, H. T. Lynch, F. S. M. Hilbers, C. J. van Asperen, H. F. A. Vasen, D. Goldgar, P. Radice, P. Devilee, and J. Benitez. Whole exome sequencing suggests much of non-brca1/brca2 familial breast cancer is due to moderate and low penetrance susceptibility alleles. *PloS One*, 8(2) :e55681, 2013. PMID : 23409019 PMCID : PMC3568132.

[318] D. J. Park, F. Lesueur, T. Nguyen-Dumont, M. Pertesi, F. Odefrey, F. Hammet, S. L. Neuhausen, E. M. John, I. L. Andrulis, M. B. Terry, M. Daly, S. Buys, F. Le Calvez-Kelm, A. Lonie, B. J. Pope, H. Tsimiklis, C. Voegele, F. M. Hilbers, N. Hoogerbrugge, A. Barroso, A. Osorio, Breast Cancer Family Registry, Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer, G. G. Giles, P. Devilee, J. Benitez, J. L. Hopper, S. V. Tavtigian, D. E. Goldgar, and M. C. Southey. Rare mutations in xrcc2 increase the risk of breast cancer. *American Journal of Human Genetics*, 90(4) :734–739, Apr. 2012. PMID : 22464251 PMCID : PMC3322233.

[319] K. Snape, E. Ruark, P. Tarpey, A. Renwick, C. Turnbull, S. Seal, A. Murray, S. Hanks, J. Douglas, M. R. Stratton, and N. Rahman. Predisposition gene identification in common cancers by exome sequencing : insights from familial breast cancer. *Breast Cancer Research and Treatment*, 134(1) :429–433, July 2012. PMID : 22527104 PMCID : PMC3781770.

[320] E. R. Thompson, M. A. Doyle, G. L. Ryland, S. M. Rowley, D. Y. H. Choong, R. W. Tothill, H. Thorne, kConFab, D. R. Barnes, J. Li, J. Ellul, G. K. Philip, Y. C. Antill, P. A. James, A. H. Trainer, G. Mitchell, and I. G. Campbell. Exome sequencing identifies rare deleterious mutations in dna repair genes fancc and blm as potential breast cancer susceptibility alleles. *PLoS genetics*, 8(9) :e1002894, Sept. 2012. PMID : 23028338 PMCID : PMC3459953.

[321] H. Wen, Y. C. Kim, C. Snyder, F. Xiao, E. A. Fleissner, D. Becirovic, J. Luo, B. Downs, S. Sherman, K. H. Cowan, H. T. Lynch, and S. M. Wang. Family-specific, novel, deleterious germline variants provide a rich resource to identify genetic predispositions for brcax familial breast cancer. *BMC cancer*, 14 :470, 2014. PMID : 24969172 PMCID : PMC4083142.

[322] F. O. Walker. Huntington's disease. *The Lancet*, 369(9557) :218–228, Jan. 2007.

[323] A. Ummat and A. Bashir. Resolving complex tandem repeats with long reads. *Bioinformatics (Oxford, England)*, July 2014. PMID : 25028725.

[324] M. Baker. Structural variation : the genome's hidden architecture. *Nature Methods*, 9(2) :133–137, Jan. 2012.

[325] B. J. Raphael. Chapter 6 : Structural variation and medical genomics. *PLoS Comput Biol*, 8(12) :e1002821, Dec. 2012.

[326] J. Wang, C. G. Mullighan, J. Easton, S. Roberts, S. L. Heatley, J. Ma, M. C. Rusch, K. Chen, C. C. Harris, L. Ding, L. Holmfeldt, D. Payne-Turner, X. Fan, L. Wei, D. Zhao,

J. C. Obenauer, C. Naeve, E. R. Mardis, R. K. Wilson, J. R. Downing, and J. Zhang. Crest maps somatic structural variation in cancer genomes with base-pair resolution. *Nature Methods*, 8(8) :652–654, June 2011.

[327] R. A. Burrell and C. Swanton. Tumour heterogeneity and the evolution of polyclonal drug resistance. *Molecular Oncology*, 2014.

[328] R. M. Cantor, K. Lange, and J. S. Sinsheimer. Prioritizing gwas results : A review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, 86(1) :6–22, Jan. 2010.

[329] N. Atias, S. Istrail, and R. Sharan. Pathway-based analysis of genomic variation data. *Current Opinion in Genetics & Development*, 23(6) :622–626, Dec. 2013. PMID : 24209906.

[330] M. D. M. Leiserson, J. V. Eldridge, S. Ramachandran, and B. J. Raphael. Network analysis of gwas data. *Current Opinion in Genetics & Development*, 23(6) :602–610, Dec. 2013. PMID : 24287332 PMCID : PMC3867794.

[331] M.-L. N. McDonald, M. Mattheisen, M. H. Cho, Y.-Y. Liu, B. Harshfield, C. P. Hersh, P. Bakke, A. Gulsvik, C. Lange, T. H. Beaty, and E. K. Silverman. Beyond gwas in copd : Probing the landscape between gene-set associations, genome-wide associations and protein-protein interaction networks. *Human Heredity*, 78(3) :131–139, Aug. 2014. PMID : 25171373.

[332] M. A. Mooney, J. T. Nigg, S. K. McWeeney, and B. Wilmot. Functional and genomic context in pathway analysis of gwas data. *Trends in genetics : TIG*, 30(9) :390–400, Sept. 2014. PMID : 25154796.

PUBLICATIONS

Publications inclues :

1. **S. Blein**, S. Berndt, A.D. Joshi, D. Campa, R.G. Ziegler, E. Riboli, D.G. Cox, on behalf of the NCI Breast and Prostate Cancer Cohort Consorium. Factors associated with oxidative stress and cancer risk in the Breast and Prostate Cancer Cohort Consortium. Free Radic. Res. (2014). doi :10.3109/10715762.2013.875168

2. A. Véron, **S. Blein** & D.G. Cox. Genome-wide association studies and the clinic : a focus on breast cancer. Biomark. Med. 8, 287–296 (2014).

3. **S. Blein**, C. Bardel, V. Danjean, L. McGuffog, S. Healey, *et al.* An original phylogenetic approach identified mitochondrial haplogroup T1a1 as inversely associated with breast cancer risk in BRCA2 mutation carriers. (*Submitted to Breast Cancer Research*)

Free Radical Research, March 2014; 48(3): 380–386

ISSN 1071-5762 print/ISSN 1029-2470 online
DOI: 10.3109/10715762.2013.875168

informa
healthcare

ORIGINAL ARTICLE

Factors associated with oxidative stress and cancer risk in the Breast and Prostate Cancer Cohort Consortium

S. Blein[1,2,3,4,5], S. Berndt[6], A. D. Joshi[7], D. Campa[8], R. G. Ziegler[6], E. Riboli[9], D. G. Cox[1,2,3,4,5,9] & on Behalf of the NCI Breast and Prostate Cancer Cohort Consortium*

[1]*Université de Lyon, Lyon, France,* [2]*Université Lyon 1, ISPB, Lyon, France,* [3]*INSERM U1052, Centre de Recherche en Cancérologie de Lyon, Lyon, France,* [4]*CNRS UMR5286, Centre de Recherche en Cancérologie de Lyon, Lyon, France,* [5]*Centre Léon Bérard, Lyon, France,* [6]*Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA,* [7]*Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA,* [8]*German Cancer Research Center (DKFZ), Heidelberg, Germany, and* [9]*Department of Epidemiology and Biostatistics, School of Public Health, Imperial College, London, UK*

Abstract

Both endogenous factors (genomic variations) and exogenous factors (environmental exposures, lifestyle) impact the balance of reactive oxygen species (ROS). Variants of the *ND3* (rs2853826; G10398A) gene of the mitochondrial genome, manganese superoxide dismutase (*MnSOD*; rs4880 Val16Ala) and glutathione peroxidase (*GPX-1*; rs1050450 Pro198Leu), are purported to have functional effects on regulation of ROS balance. In this study, we examined associations of breast and prostate cancer risks and survival with these variants, and interactions between rs4880–rs1050450, and alcohol consumption–rs2853826. Nested case-control studies were conducted in the Breast and Prostate Cancer Cohort Consortium (BPC3), consisting of nine cohorts. The analyses included over 10726 post-menopausal breast and 7532 prostate cancer cases with matched controls. Logistic regression models were used to evaluate associations with risk, and proportional hazard models were used for survival outcomes. We did not observe significant interactions between polymorphisms in *MnSOD* and *GPX-1*, or between mitochondrial polymorphisms and alcohol intake and risk of either breast (p-interaction of 0.34 and 0.98, respectively) or prostate cancer (p-interaction of 0.49 and 0.50, respectively). We observed a weak inverse association between prostate cancer risk and GPX-1 Leu198Leu carriers (OR 0.87, 95% CI 0.79–0.97, $p = 0.01$). Overall survival among women with breast cancer was inversely associated with G10398 carriers who consumed alcohol (HR 0.66 95% CI 0.49–0.88). Given the high power in our study, it is unlikely that interactions tested have more than moderate effects on breast or prostate cancer risk. Observed associations need both further epidemiological and biological confirmation.

Keywords: *MnSOD*, *GPX-1*, mitochondria, alcohol, breast, prostate

Introduction

Reactive oxygen species (ROS) are naturally occurring chemical entities derived from the metabolism of a number of compounds, in addition to being produced during oxidative respiration in the mitochondria. ROS are not only cytotoxic but also mutagenic. Elevated levels of ROS and down regulation of ROS scavengers and/or antioxidant enzymes, which can lead to oxidative stress, are associated with a number of human diseases including various cancers [1,2]. Oxidative damage to DNA can lead to changes such as single base modifications, gene duplications and the activation of oncogenes, and may be involved in the initiation of cancer.

Manganese superoxide dismutase (MnSOD) and glutathione peroxidase (GPX) are antioxidant enzymes, coded by the *MnSOD* and *GPX-1* genes, respectively. The toxic superoxide anion ($O_2^{\cdot -}$) is a naturally occurring by-product

*Co-Authors from the BPC3: M. M. Gaudet, V. L. Stevens, W. R. Diver, S. M. Gapstur (American Cancer Society); S. J. Chanock, R. N. Hoover, M. Yeager, D. Albanes (CNRS UMR5286, Centre de Recherche en Canc é rologie de Lyon); J. Virtamo (National Institute for Health and Welfar); E. D. Crawford (University of Colorado AMC Campus); C. Isaacs (Comprehensive Cancer Center, Georgetown University Medical Center); C. Berg (National Cancer Institute); D. Trichopoulos (Harvard School of Public Health, Bureau of Epidemiologic Research, Academy of Athens); S. Panico (Federico II University); P. H. Peeters (Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht); M. Johansson (International Agency for Research on Cancer, Umea University); K. T. Khaw (Clinical Gerontology Unit, School of Clinical Medicine, University of Cambridge); P. Kraft, D. J. Hunter, S. Lindström, J. Ma, M. Stampfer (Harvard School of Public Health); J. M. Gaziano (Brigham and Women's Hospital); E. Giovannucci, W. H. Willett (Harvard School of Public Health); S. E. Hankinson (Brigham and Women's Hospital, Keck School of Medicine and Norris Comprehensive Cancer Center, University of Southern California); I. M. Lee, J. Buring (Harvard School of Public Health, Brigham and Women's Hospital); B. Henderson (Keck School of Medicine and Norris Comprehensive Cancer Center, University of Southern California); L. L. Marchand, L. Kolonel (Epidemiology Program, University of Hawaii Cancer Center); C. J. Haiman (Keck School of Medicine and Norris Comprehensive Cancer Center, University of Southern California).

Correspondence: David G. Cox, Centre Léon Bérard, 28 rue Laënnec, 69373 Lyon Cedex 8, France. Tel: (+33) 4-78-78-59-12. E-mail: david.cox@inserm.fr

(Received date: 18 October 2013; Accepted date: 10 December 2013; Published online: 10 January 2014)

of the mitochondrial electron transport chain. MnSOD acts as the first protective barrier against superoxide anion by metabolizing it into hydrogen peroxide. Hydrogen peroxide is further detoxified by *GPX-1*, a selenium-containing protein acting as a detoxifier of hydrogen products. These two enzymes contribute to regulate ROS exposures in the cell, and to prevent oxidative stress.

Various exogenous agents are known to affect ROS balance. Ozone and cigarette smoke [3], absorption of heavy metal particles [4], high dietary fat intake [5], and alcohol consumption [6–8] are environmental and lifestyle factors contributing to increased ROS production. Endogenous factors, and in particular genomic variations, may also play a key role in the regulation of ROS balance. Variants in genes involved in ROS regulation (antioxidant enzymes, ROS scavengers) or production (mitochondrial genes in charge of the respiratory chain) have been shown to alter their function and efficiency [9,10].

For complex diseases like breast or prostate cancer, several low penetrance polymorphisms could have a synergistic effect leading to higher cancer risk [11]. Both breast [12,13] and prostate [14,15] cancer are reported to be linked to oxidative stress. Studies examining this hypothesis in specific candidate genes have been conducted in the Nurses' Health Study [16–18] with particular focus on gene-by-gene and gene-by-environment interactions related to ROS exposure.

The underlying hypothesis is that variations in the coding sequence leading to amino acid changes in the synthesized enzymes alter their function and efficiency [19–21], contributing to a deregulation of the balance of ROS, ultimately leading to oxidative stress. Individuals homozygous for the Ala16 allele of MnSOD and Leu198 allele of GPX-1 were observed to have a 1.87 fold increase in breast cancer risk compared with Val16 and Pro198 carriers, with a p-value for interaction of 0.03 in the Nurses' Health Study [16]. Furthermore, recently Méplan et al. showed an association between breast cancer risk and alternative allele of GPX-1 polymorphism Pro198Leu ($p = 0.027$) [23].

A second investigation in the Nurses' Health Study explored the impact of alcohol consumption on breast cancer risk for individuals carrying the mitochondrial SNP rs2853826/A10398G in the *ND3* gene [17]. Variations in the mitochondrial genome could affect the processing and the efficiency of the mitochondrial electron transport chain, leading to an increase in ROS production, and thus be associated with an increased risk of breast cancer. We previously showed that the 10398G allele modifies the association between alcohol consumption and breast cancer risk, with an odds ratio of 1.52 for drinkers compared with non-drinkers. No association between alcohol consumption and BC risk was observed among women carrying the 10398A allele.

In this study, we aimed to further investigate these hypotheses in independent populations. In addition, because oxidative stress may be linked to both breast and prostate cancers, we extend our research to prostate cancer risk. We genotyped rs2853826, rs4880, and rs1050450 in the NCI Breast and Prostate Cancer Cohort Consortium (BPC3), a large international consortium combining resources of nine well-established cohort studies [22]. While a large number of subjects in these cohorts were included in genome wide association scans (GWAS [24,25]), these three polymorphisms were not well represented on the specific products used in previous analyses. As such, any attempt to detect gene-by-gene and gene-by-environment interactions in large-scale genotyping efforts would not be informative for these specific hypotheses.

Materials and methods

BPC3

The National Cancer Institute BPC3 has been described previously [22]. In brief, the consortium combines resources from nine well-established cohort studies: the Alpha-Tocopherol, Beta-Carotene Cancer Prevention, American Cancer Society Cancer Prevention Study II, the European Prospective Investigation into Cancer and Nutrition Cohort (EPIC–composed of cohorts from Denmark, France, Great Britain, Germany, Greece, Italy, the Netherlands, Spain, and Sweden), the Health Professionals Follow-up Study, the Multiethnic Cohort, the Physicians' Health Study, the Nurses' Health Study (NHS), the Women's Health Study (WHS), and the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. Each study was approved by local Institutional Review Boards, and informed consent was obtained from all subjects.

Each cohort has its own method of case ascertainment, exposure assessment, and matching criteria to health controls [22]. A total of 13511 post-menopausal women diagnosed with BC, and 8490 men diagnosed with prostate cancer were included in our study to analyze the interaction between the two polymorphisms, respectively, in *MnSOD* and *GPX-1*. A total of 10726 post-menopausal women diagnosed with BC, and 7532 men diagnosed with prostate cancer were included in analyzing the interaction between mitochondrial polymorphism rs2853826/A10398G and alcohol consumption. As described in Hendrickson et al. [26], ER/PR status is known for around 60–80% of breast cancer cases participating in the BPC3. ER status among these cases follows the general distribution of around 20% ER− and 80% ER+ breast cancer. Given that a large number of cancer cases were ascertained prior to routine use, HER2 amplification and other biomarkers such as EGFR expression are not available.

Genotyping

Three single nucleotide polymorphisms (SNPs) were genotyped: rs1050450 (*GPX-1*), rs4880 (*MnSOD*), and the mitochondrial SNP rs2853826. For one cohort (PLCO) the SNP rs8031 (*MnSOD*), which is in high linkage disequilibrium with rs4880 ($r^2 = 0.95$ based on

Hapmap genotype frequencies) was genotyped to infer the genotype of rs4880. Genotyping was performed via TaqMan assays. rs4880 and rs1050450 were in Hardy–Weinberg Equilibrium among controls, stratified by cohort. A total of 334 subjects were excluded due to poor genotyping. For each SNP and each cohort, call rates were greater than 0.91. Cohorts with less than 90% success for a given SNP were removed from analyses for that SNP.

Statistical analysis

Regression analyses were carried out in R. Logistic regressions were conducted for each type of cancer, and for each of the two interactions tested. All analyses were performed under a recessive genetic model testing homozygotes for Leu allele versus Pro allele carriers, as previously published [17,16]. Adjustment factors for each analysis are summarized in Table I. Wald and Likelihood Ratio Test p-values were computed to assess statistical interactions. Power calculations were performed using the Quanto software [26] (see Supplementary Table I to be found online at http://informahealthcare.com/doi/abs/10.3109/10715762.2013.875168). Survival analyses among cases with respect to genotype and lifestyle factors were performed using proportional hazard models with the *R-package* survival. Heterogeneity tests were performed to evaluate the similarity of associations between cohorts, and random effects model estimates were retained when heterogeneity tests were significant ($p < 0.05$). A meta-analysis combining our data with published data concerning GPX-1 rs1050450 and PC was performed. Six studies were initially selected to be included [27–32]. However, Steinbrecher et al. [27] used data from the EPIC cohort, which is part of the BPC3. Thus we excluded this study, which might not be independent of our study. Results from Cheng et al. [29] were also excluded because information regarding homozygous and heterozygous Leu allele carriers was not presented. Dixon and Grubbs tests, from R package *outliers*, were used to assess whether outliers were present among all considered studies.

Results

For prostate cancer, we found an inverse association among homozygotes for the variant allele at rs1050450 of *GPX-1* (OR 0.87, 95% [0.79–0.97], Table I). Figure 1 presents results of a meta-analysis pooling our study of rs1050450 and prostate cancer with other published studies [28,30–32]. We observed heterogeneity between studies ($p = 0.0033$), and random effects model estimation of global odds-ratio and confidence interval is 1.19 [0.79–1.80]. The Dixon and Grubbs tests identified the study of Kucukgergin et al. [31] as an outlier, and meta-analysis after exclusion of this study showed no between-study heterogeneity, with a fixed effects model odds-ratio of 0.90 (95% Confidence Interval 0.81–0.99). No interaction was detected between alcohol consumption and the *ND3* A10398G polymorphism (p-interaction = 0.50) with respect to prostate cancer risk or survival.

For breast cancer, we did not observe associations for the interaction between rs4880 Val16Ala in *MnSOD* and rs1050450 Pro198Leu in *GPX-1* (Table I). Our

Table I. Associations with breast and prostate cancer.

	Genotype	Cases (%)	Controls (%)	OR (95% CI)
Interaction between GPX-1 Pro198Leu and MnSOD Val16Ala and breast cancer risk[a]	Pro198 carrier and Val16 carrier	3215 (66.3)	3685 (65.9)	1 (Ref.)
	Pro198 carrier and Ala16Ala	1134 (23.4)	1326 (23.7)	1.00 (0.96–1.03)
	Leu198Leu and Val16 carrier	371 (7.6)	442 (7.9)	1.00 (0.97–1.02)
	Leu198Leu and Ala16Ala	132 (2.7)	139 (2.5)	1.03 (0.97–1.09)
Interaction between mitochondrial A10398G and alcohol consumption on breast cancer risk[b]	A10398, non-Drinkers	1114 (33.7)	1443 (36.2)	1 (Ref.)
	A10398, Drinkers	1507 (45.6)	1732 (43.5)	1.13 (1.02–1.26)
	G10398, non-Drinkers	294 (8.9)	372 (9.3)	1.03 (0.87–1.23)
	G10398, Drinkers	391 (11.8)	436 (10.9)	1.16 (0.99–1.36)
Association between GPX-1 Pro198Leu and prostate cancer risk[c]	Pro198 Carrier	6688 (89.4)	6510 (88.2)	1 (Ref.)
	Leu198Leu	792 (10.6)	867 (11.8)	0.87 (0.79–0.97)
Interaction between GPX-1 Pro198Leu and MnSOD Val16Ala and prostate cancer risk[d]	Pro198 carrier and Val16 carrier	4223 (66.2)	4230 (66.3)	1 (Ref.)
	Pro198 carrier and Ala16Ala	1473 (23.1)	1396 (21.9)	1.06 (0.97–1.15)
	Leu198Leu and Val16 carrier	507 (7.9)	573 (89)	0.88 (0.76–0.98)
	Leu198Leu and Ala16Ala	176 (2.8)	208 (3.3)	0.84 (0.67–1.02)
Interaction between mitochondrial A10398G and alcohol consumption on prostate cancer risk[e]	A10398, non-Drinkers	389 (10.6)	429 (11.1)	1 (Ref.)
	A10398, Drinkers	2448 (66.7)	2596 (67.2)	1.15 (0.99–1.33)
	G10398, non-Drinkers	135 (3.7)	131 (3.4)	1.12 (0.85–1.48)
	G10398, Drinkers	698 (19.0)	706 (18.3)	1.16 (0.97–1.38)

[a]P-interaction = 0.34. Data restricted to post-menopausal women. Unconditional logistic regression controlled for age at blood draw, age at menarche, age at menopause, body–mass index, family history of breast cancer, and cohort.
[b]P-interaction = 0.98. Data restricted to post-menopausal women. Unconditional logistic regression controlled for age at blood draw, age at menarche, age at menopause, body–mass index, family history of breast cancer, and cohort.
[c]Unconditional logistic regression controlled for age at diagnosis, alcohol consumption, and cohort.
[d]P-interaction = 0.44. Unconditional logistic regression controlled for age at diagnosis, alcohol consumption, and cohort.
[e]P-interaction = 0.50. Unconditional logistic regression controlled for age at diagnosis and cohort.

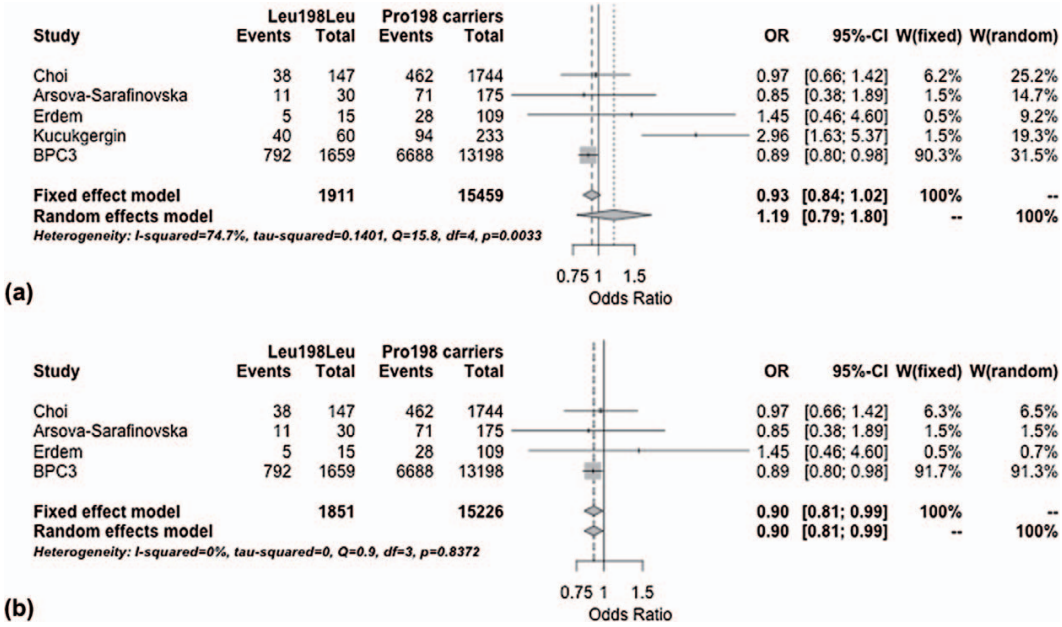


Figure 1. Meta-analysis pooling already published studies with our study in the BPC3, with and without study from Kucukgergin et al. [31]. (a) *Top.* Forest-plots of odds-ratios with 95% confidence intervals for GPX-1 polymorphism and risk of prostate cancer Leu198Leu vs. Pro carriers (Ref.) (b) *Bottom.* Forest-plots of odds-ratios with 95% confidence intervals for GPX-1 polymorphism and risk of prostate cancer—Leu198Leu versus Pro carriers (Ref.)—after exclusion of Kucukgergin et al. [31].

study had greater than 95% power to detect an odds ratio of 1.87 as found in previous studies, at an α of 0.05, for an interaction between the recessive model for two polymorphisms where neither polymorphism alone is associated with risk. Neither of the two variants tested had an independent association with breast cancer risk. Carriers of both variant alleles were equally likely to develop breast cancer compared with reference allele carriers (OR = 1.03, 95% CI [0.97–1.09]). No statistical interaction was detected between rs1050540 and rs4880 on breast cancer risk (p-interaction = 0.34). No difference in association between alcohol consumption and breast cancer risk was observed among carriers of the G or A alleles of the *ND3* A10398G polymorphism (p-interaction = 0.98). All results for breast cancer were similar when the NHS and WHS were removed (results not shown). Survival curves (Figure 2 and Table II) are not statistically different between groups defined by genotypes for all analyses except for the mitochondrial A10398G genotype and alcohol consumption status among breast cancer patients. We observed a substantial difference between overall survival curves between different groups of genotype and alcohol intake (Figure 2a and b). This effect is not present for breast cancer-specific survival (Figure 2c).

Discussion

The objective of the present study was to further evaluate associations of genomic variations involved in the altera-

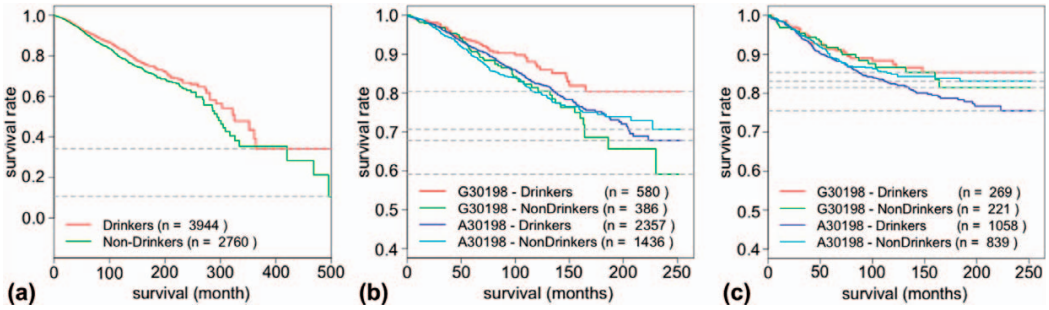


Figure 2. Survival plot (BC) with respect to genotype and alcohol consumption class. (a) Overall survival stratified (b) Overall survival stratified by alcohol consumption class and genotype (c) BC-Specific survival stratified by alcohol and genotype.

Table II. Survival analyses for breast and prostate cancer.

Cancer type	Analysis	Category	Overall survival		Specific survival	
			Odd-ratios and 95% confidence interval	Log-rank test P value	Odd-ratios and 95% confidence interval	Log-rank test P value
BREAST	2 SNPs GPX1/MnSOD*	Pro198 carrier and Val16 carrier	1 (Ref.)	0.747	1 (Ref.)	0.508
		Pro198 carrier and Ala16Ala	0.93 (0.79–1.10)		1.09 (0.87–1.36)	
		Leu198Leu and Val16 carrier	0.93 (0.72–1.21)		0.85 (0.57–1.26)	
		Leu198Leu and Ala16Ala	0.85 (0.55–1.32)		0.74 (0.38–1.43)	
	Alcohol × mtSNP A10398G*	A10398—Non-Drinkers	1 (Ref.)	0.029	1 (Ref.)	0.103
		A10398—Drinkers	0.9 (0.76–1.07)		1.24 (0.97–1.57)	
		G10398—Non-Drinkers	1.03 (0.77–1.36)		0.95 (0.62–1.44)	
		G10398—Drinkers	0.66 (0.49–0.88)		0.84 (0.57–1.26)	
	Alcohol effect Only*	Non-Drinkers	1 (Ref.)	0.003	1 (Ref.)	0.104
		Drinkers	0.84 (0.74–0.94)		1.15 (0.97–1.36)	
PROSTATE	2 SNPs GPX1/MnSOD**	Pro198 carrier and Val16 carrier	1 (Ref.)	0.855	1 (Ref.)	0.356
		Pro198 carrier and Ala16Ala	1.00 (0.89–1.12)		0.93 (0.77–1.14)	
		Leu198Leu and Val16 carrier	1.00 (0.84–1.18)		0.75 (0.54–1.04)	
		Leu198Leu and Ala16Ala	1.13 (0.86–1.49)		1.05 (0.64–1.7)	
	Alcohol × mtSNP A10398G**	A10398—Non-Drinkers	1 (Ref.)	0.558	1 (Ref.)	0.148
		A10398—Drinkers	1.07 (0.87–1.32)		1.17 (0.82–1.67)	
		G10398—Non-Drinkers	0.85 (0.57–1.27)		0.49 (0.21–1.16)	
		G10398—Drinkers	1.09 (0.86–1.38)		1.13 (0.75–1.69)	

*Adjusted for Cohort and Age at Breast Cancer diagnosis.
**Adjusted for Cohort and Age at Prostate Cancer diagnosis.

tion of ROS balance with respect to breast and prostate cancer risks and survival. First, we focused on variations occurring in *MnSOD* and *GPX-1*, two genes encoding antioxidant enzymes that protect cellular DNA from oxidative damage. Carriers of both variant alleles for rs4880 Val16Ala in *MnSOD* and rs1050450 Pro198Leu in *GPX-1* did not have a change in risk of developing breast or prostate cancer.

Interestingly, we observed a weak inverse association between being homozygous Leu198Leu for rs1050450 (*GPX-1*) and risk of prostate cancer. This polymorphism has not been represented on previous GWAS products, and was only recently genotyped in the HapMap or in the 1000 Genomes project, so it is possible that this putative association was not detectable in prior GWAS, and inconsistent results were found for this association in candidate analyses. Although some studies observed a trend toward an inverse association between rs1050450 and prostate cancer for carriers of alternative allele of rs1050450 [27], others found an absence of association [32,28,33] or in an increase in risk for carriers of alternative alleles [31]. Our study is the first that shows an inverse association with a $p < 0.05$, but also the first having statistical power to detect a weak association, and the results of the meta-analysis reinforced our observations.

With respect to breast cancer, we observed no interaction between mitochondrial SNP rs2853826 and alcohol consumption. However, women who carry the G10398 allele and consume alcohol had longer survival both in terms of overall and disease-specific survival. Conversely, women who carry the A10398 allele and consume alcohol had reduced survival as compared with those who carry the same allele and do not consume alcohol. One hypothesized explanation for these results is the inverse association between moderate alcohol consumption and cardiovascular disease. As the majority of women in this study are post-menopausal (72%), and follow-up time is long (on average 8 years after diagnosis), there is potential for heart disease, independent of breast cancer status, to be a competing outcome. Furthermore, a number of adjuvant breast cancer therapies such as bevacizumab, taxane-anthracycline, and trastuzumab are associated with cardiovascular complications during BC treatment. Given the need to simplify our definition of alcohol consumption in order to combine data across participating cohorts, we have used a crude ever/never categorization. Further clarification of this putative association by more careful classification of alcohol consumption, particularly by type and quantity, may shed further light on these results. Additionally, information on treatments, unavailable for the vast majority of our study participants, will provide additional clarification in survival analyses.

Conclusion

In conclusion, we did not observe interactions between rs4880 in *MnSOD* and rs1050450 in *GPX-1* or rs2853826 and alcohol consumption and risk of breast or prostate cancer. However we did observe a putative inverse association between rs1050450 in *GPX-1* and prostate cancer risk, and a novel interaction between alcohol consumption and rs2853826 in the mitochondrial *NAD3* gene on breast cancer survival.

Acknowledgments

The authors thank the CPS-II participants and Study Management Group for their invaluable contributions to this research. The authors would also like to acknowledge the contribution to this study from central cancer registries supported through the Centers for Disease Control and Prevention National Program of Cancer Registries, and cancer registries supported by the National Cancer Institute Surveillance Epidemiology and End Results program. David G. Cox is the recipient of a grant from the French Ligue Contre le Cancer, Comité de Savoie. Sophie Blein is the recipient of a CIFRE fellowship from the French ANRT and the LYRIC program.

Declaration of interest

The authors report no declarations of interest. The authors alone are responsible for the content and writing of the paper.

This work was funded by the U.S. National Institutes of Health, National Cancer Institute (cooperative agreements U01-CA98233 to David J. Hunter, U01-CA98710 to Michael J. Thun, U01-CA98216 to Elio Riboli and Rudolf Kaaks, and U01-CA98758 to Brian E. Henderson, and Intramural Research Program of NIH/NCI, Division of Cancer Epidemiology and Genetics). The American Cancer Society (ACS) funds the creation, maintenance, and updation of the Cancer Prevention Study-II (CPS-II) cohort.

References

[1] Loft S, Poulsen HE. Cancer risk and oxidative DNA damage in man. J Mol Med (Berl) 1996;74:297–312.
[2] Loft S, Høgh Danielsen P, Mikkelsen L, Risom L, Forchhammer L, Møller P. Biomarkers of oxidative damage to DNA and repair. Biochem Soc Trans 2008;36:1071–1076.
[3] Churg A. Interactions of exogenous or evoked agents and particles: the role of reactive oxygen species. Free Radic Biol Med 2003;34:1230–1235.
[4] Ercal N, Gurer-Orhan H, Aykin-Burns N. Toxic metals and oxidative stress part i: mechanisms involved in metal-induced oxidative damage. Curr Top Med Chem 2001;1:529–539.
[5] Lim S, Won H, Kim Y, Jang M, Jyothi KR, Kim Y, Dandona P, Ha J, Kim SS. Antioxidant enzymes induced by repeated intake of excess energy in the form of high-fat, high-carbohydrate meals are not sufficient to block oxidative stress in healthy lean individuals. Br J Nutr 2011;106:1544–1551.
[6] Poeschl G, Seitz HK. Alcohol and cancer. Alcohol Alcohol 2004;39:155–165.
[7] Seitz HK, Stickel F. Molecular mechanisms of alcohol-mediated carcinogenesis. Nat Rev Cancer 2007;7:599–612.
[8] Coronado GD, Beasley J, Livaudais. J Alcohol consumption and the risk of breast cancer. Salud Publica Mex 2011;53:440–447.
[9] Bai RK, Leal SM, Covarrubias D, Liu A, Wong LJC. Mitochondrial genetic background modifies breast cancer risk. Cancer Res 2007;67:4687–4694.
[10] Wang S, Wang F, Shi X, Dai J, Peng Y, Guo X, Wang X, Shen H, Hu Z. Association between manganese superoxide dismutase (MnSOD) val-9Ala polymorphism and cancer risk - a meta-analysis. Eur J Cancer 2009;45:2874–2881.
[11] Cerne JZ, Pohar-Perme M, Novakovic S, Frkovic-Grazio S, Stegel V, Gersak K. Combined effect of CYP1B1, COMT, GSTP1, and MnSOD genotypes and risk of postmenopausal breast cancer. J Gynecol Oncol 2011;22:110.
[12] Sener DE, Gönenc A, Akinci M, Torun M. Lipid peroxidation and total antioxidant status in patients with breast cancer. Cell Biochem Funct 2007;25:377–382. PMID: 16447143.
[13] Kang DH. Oxidative stress, DNA damage, and breast cancer. AACN Clin Issues 2002;13:540–549. PMID: 12473916.
[14] Gupta-Elera G, Garrett AR, Robison RA, O'Neill KL. The role of oxidative stress in prostate cancer. Eur J Cancer Prev 2012;21:155–162. PMID: 21857523.
[15] Khandrika L, Kumar B, Koul S, Maroni P, Koul HK. Oxidative stress in prostate cancer. Cancer Lett 2009;282:125–136.
[16] Cox DG, Tamimi RM, Hunter DJ. Gene x gene interaction between MnSOD and GPX-1 and breast cancer risk: a nested case-control study. BMC Cancer 2006;6:217.
[17] Pezzotti A, Kraft P, Hankinson SE, Hunter DJ, Buring J, Cox DG. The mitochondrial A10398G polymorphism, interaction with alcohol consumption, and breast cancer risk. PLoS One 2009;4:e5356.
[18] Ibarrola-Villava M, Peña-Chilet M, Fernandez LP, Aviles JA, Mayor M, Martin-Gonzalez M, et al. Genetic polymorphisms in DNA repair and oxidative stress pathways associated with malignant melanoma susceptibility. Eur J Cancer 2011;47:2618–2625.
[19] Pietras T, Szemraj J, Witusik A, Hołub M, Panek M, Wujcik R, Górski P. The sequence polymorphism of MnSOD gene in subjects with respiratory insufficiency in COPD. Med Sci Monit 2010;16:CR427–CR432.
[20] Chen Y, Pei J. Possible risk modifications in the association between MnSOD Ala9Val polymorphism and breast cancer risk: subgroup analysis and evidence-based sample size calculation for a future trial. Breast Cancer Res Treat 2011;125:495–504.
[21] Ezzikouri S, El Feydi AE, Afifi R, Benazzouz M, Hassar M, Pineau P, Benjelloun S. Poly- morphisms in antioxidant defence genes and susceptibility to hepatocellular carcinoma in a moroccan population. Free Radic Res 2010;44:208–216.
[22] Hunter DJ, Riboli E, Haiman CA, Albanes D, Altshuler D, Chanock SJ, et al. A candidate gene approach to searching for low-penetrance breast and prostate cancer genes. Nat Rev Cancer 2005;5:977–985.
[23] Méplan C, Dragsted LO, Ravn-Haren G, Tjønneland A, Vogel U, Hesketh J. Association between Polymorphisms in Glutathione Peroxidase and Selenoprotein P Genes, Glutathione Peroxidase Activity, HRT Use and Breast Cancer Risk. PLoS one 2013;8:e73316.
[24] Gudmundsson J, Sulem P, Rafnar T, Bergthorsson JT, Manolescu A, Gudbjartsson D, et al. Common sequence variants on 2p15 and xp11.22 confer susceptibility to prostate cancer. Nat Genet 2008;40:281–283.
[25] Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 2007;39:645–649.
[26] Hendrickson SJ, Lindström S, Eliassen AH, Rosner BA, Chen C, Barrdahl M, et al. Plasma carotenoid- and retinol-weighted multi-SNP scores and risk of breast cancer in the National Cancer Institute Breast and Prostate Cancer Cohort Consortium. Cancer Epidemiol Biomarkers Prev 2013;22:927–936. PMID: 23515144.
[27] Steinbrecher A, Méplan C, Hesketh J, Schomburg L, Endermann T, Jansen E, et al. Effects of selenium status and polymorphisms in selenoprotein genes on prostate cancer risk in a prospective study of european men. Cancer Epidemiol Biomarkers Prev 2010;19:2958–2968. PMID: 20852007.
[28] Erdem O, Eken A, Akay C, Arsova-Sarafinovska Z, Matevska N, Suturkova L, et al. Association of GPX1

polymorphism, GPX activity and prostate cancer risk. Hum Exp Toxicol 2012;31:24–31. PMID: 21636625.
[29] Cheng TYD, Barnett MJ, Kristal AR, Ambrosone CB, King IB, Thornquist MD, et al. Genetic variation in myeloperoxidase modifies the association of serum α-tocopherol with aggressive prostate cancer among current smokers. J Nutr 2011;141:1731–1737. PMID: 21795425.
[30] Arsova-Sarafinovska Z, Matevska N, Eken A, Petrovski D, Banev S, Dzikova S, et al. Glutathione peroxidase 1 (GPX1) ge- netic polymorphism, erythrocyte GPX activity, and prostate cancer risk. Int Urol Nephrol 2009;41:63–70. PMID: 18563616.
[31] Kucukgergin C, Gokpinar M, Sanli O, Tefik T, Oktar T, Seckin. S Association between genetic variants in glutathione peroxidase 1 (GPx1) gene, GPx activity and the risk of prostate cancer. Minerva Urol Nefrol 2011;63:183–190. PMID: 21993316.
[32] Choi JY, Neuhouser ML, Barnett M, Hudson M, Kristal AR, Thornquist M, et al. Polymorphisms in oxidative Stress– Related genes are not associated with prostate cancer risk in heavy smokers. Cancer Epidemiol Biomarkers Prev 2007; 16:1115–1120.
[33] Men T, Zhang X, Yang J, Shen B, Li X, Chen D, Wang J. The rs1050450 C > T polymorphism of GPX1 is associated with the risk of bladder but not prostate cancer: evidence from a meta-analysis. Tumor Biol 2013; PMID: 23975365.

Supplementary material available online

Supplementary Table I.

Review

Genome-wide association studies and the clinic: a focus on breast cancer

Breast cancer is the most frequently diagnosed cancer among women worldwide, and has long been considered to be a genetic disease. A wide range of genetic variants, both rare mutations and more common variants, have been shown to influence breast cancer risk. In particular, recent studies have identified a number of common genetic variants, or single nucleotide polymorphisms, that are associated with breast cancer risk. In this review, we will briefly present the genetic epidemiology of breast cancer, genome-wide association study technology and how this technology may influence breast cancer screening in the clinic.

KEYWORDS: breast cancer ▪ genome-wide association study ▪ GWAS ▪ screening ▪ single nucleotide polymorphism ▪ SNP ▪ susceptibility

Amélie Véron[1,2,3,4,5], Sophie Blein[1,2,3,4,5] & David G Cox*[1,2,3,4,5]

[1]*Université de Lyon, F-69000 Lyon, France*
[2]*Université Lyon 1, ISPB, Lyon, F-69622, France*
[3]*INSERM U1052, Centre de Recherche en Cancérologie de Lyon, F-69000 Lyon, France*
[4]*CNRS UMR5286, Centre de Recherche en Cancérologie de Lyon, F-69000 Lyon, France*
[5]*Centre Léon Bérard, F-69008 Lyon, France*
**Author for correspondence: david.cox@lyon.unicancer.fr*

Epidemiology of breast cancer

Breast cancer (BC) is a common cause of cancer death among women, both in developed and developing countries, and BC incidence is steadily increasing worldwide. More than 1.3 million people were diagnosed with BC in 2008, representing 20.3% of all cancers in women, and approximately 10.9% of cancers in men and women combined [1]. The lowest incidence rate (<40/100,000 women) is observed in developing countries whereas the highest incidence rates (>80/100,000 women) are observed in developed countries. However, the relatively low incidence rate observed in developing countries is expected to increase significantly during the next few decades [2].

BC is a complex and heterogeneous multifactorial disease with both environmental and genetic risk factors. Based on American Cancer Society reports between 2009 and 2012 [3–5], substantial variations in BC risk are observed between ethnicities in the USA. White women have overall the highest BC risk (age-adjusted rate of 125.4 cases per 100,000 women). However, black women aged younger than 45 years have a slightly higher BC risk than white women of the same age, and are more likely to be diagnosed with larger tumors. Hispanic/Latina women also have lower incidence (age-adjusted rate of 91.0 cases per 100,000 women) of BC than white women, but similar to black women they are more likely to be diagnosed with large, advanced tumors. Furthermore, genetic predisposing features have already been identified in specific ethnic groups [6,7]. Nongenetic risk factors, such as environment and lifestyle, also have a major impact on BC risk. Despite the fact that incidence rates have historically been four- to seven-times higher in the USA than in China or Japan, BC risk for female immigrants from Asia rise to reach American incidence rates in only a few generations [8]. Furthermore, Asian immigrants who have lived in the USA for at least 10 years have an 80% increased risk of BC compared with women in their country of origin. Many environmental, lifestyle and nongenetic risk factors have been associated with an increase in incidence of BC. Alcohol intake, smoking, oral contraception, age, mammographic density and hormone replacement therapy are established risk factors (Table 1), but this list is not exhaustive, and others factors such as radiations exposure, levels of circulating cholesterol or diet generally have also been proposed.

Genetic variants & BC risk

One of the strongest risks factors for BC is the existence of a family history of the disease [9]. The risk for a woman to develop BC increases with the number of relatives diagnosed with this disease. However, only 5–10% of all BCs are attributed to strong inherited components, with approximately 4–5% due to high-penetrance predisposing genes [10–12].

Currently, three high-penetrance genes have been identified: *BRCA1*, *BRCA2* and *TP53*. The pooled frequency of mutations in these genes in the general population is estimated at 0.4% [13,14] and, therefore, despite their high penetrance, they explain only a small percentage of the genetic risk of BC. *BRCA1* was cloned in the early 1990s, and individuals carrying mutations

Future Medicine part of fsg

Table 1. Modifiable and others risk factors for breast cancer.

Risk factor	Effect	Modifiers/complexity	Ref.
Smoking	↑	Age at initiation Initiation before	[81]
Alcohol	↑	Tumor subtype	[82,83]
Hormone replacement therapy	↑	Menopausal status Duration of treatment Pre-existing lesions	[84] [42] [43]
Oral contraception	↓	Duration of use Time since cessation Duration before First full-term pregnancy	[85] [86] [87] [88]
BMI	↑↓	Menopausal status	[89]
Age	↑	None	[3]
Mammographic density	↑	Age, BMI and parity Hormonal therapy	[90] [91]

↑: Increase; ↓: Decrease.

of *BRCA1* account for approximately 7–10% of BC cases. Pathogenic mutations in *BRCA1* confer a lifetime risk of BC of 60–85% [15,16]. *BRCA2* mutations account for approximately 10% of families with BCs [17]. Mutations in *BRCA2* confer a lifetime risk of approximately 40–85% [15,16]. Germline mutations in another gene, *TP53*, are rare, but cause a form of very early-onset BC. A total of 30% of female carriers of mutations of this gene develop BC before the age of 30 years. Mutations of *TP53* confer an 18–60-fold increase in risk of BC for women <45 years compared with the general population. Other genes such as *PTEN* [18], *STK11* [19] or *CDH1* [20] have been associated with a high lifetime risk (~40–60%) of BC [21].

Four genes have been identified as conferring a moderate (two- to four-fold) BC risk. Mutations in these genes remain rare, with a population frequency of <0.4%. For example, studies show that carriers of heterozygous mutations on *ATM* have a approximately a twofold increase (odds ratio [OR]: 2.23; 95% CI: 1.16–4.28) of BC risk for women under the age of 50 years [22]. The penetrance of mutations in *ATM* is approximately 15%, and it is currently not possible to determine which mutation carriers will develop BC. A particular mutation of *CHEK2*, 1100delC, increases BC risk approximately twofold [23]. Truncating mutations in the BRCA1 partner *BRIP1* were identified in BC families, with a twofold (95% CI: 1.2–3.2) increase in BC risk [24]. Finally, for *PALB2*, a partner of BRCA2 [25], the relative risk of mutations in this gene associated with respect to BC is 2.3 (95% CI: 1.4–3.9) [9].

Prior to the genome-wide association study (GWAS) era, candidate studies were the most efficient at evaluating the association between genetic polymorphisms and BC risk. Early studies were based on genotyping candidate single polymorphisms with *a priori* functional consequences, often in genes thought to be highly influential with respect to cancer risk. The combination of increasing capacity to genotype by PCR-based methods and knowledge of the correlation structure of the genome led to using haplotype-based approaches in candidate genes to evaluate association of risk with disease. Progression to chip-based methods further increased genotyping efficiency and led to the development of candidate pathway approaches.

Unfortunately, despite decades of effort, the candidate single nucleotide polymorphism (SNP) approach met with limited success with respect to the detection of polymorphisms associated with BC risk. A paper published by the Breast Cancer Association Consortium (BCAC) examined the association between the nine most promising candidate SNPs combining data from numerous studies [26]. Depending on the SNP, data from 11,391–18,290 cases and 14,753–22,670 controls were available to evaluate the associations. Of these polymorphisms, only the D302H (rs1045485) variant in *CASP8* and the L10P (rs1982073) variant in *TGFB1* were shown to be associated with BC risk. A study published at approximately the same time found the opposite association between *TGFB1* L10P and BC risk, and upon combining analyses from both studies, no association was observed. This leaves only *CASP8* D302H as a candidate variant associated with a moderate change in BC risk [27].

In a similar fashion, candidate gene and pathway approaches have not been overly successful in identifying genetic variants that are associated with BC risk. The Breast and Prostate Cancer Cohort Consortium (BPC3) applied this approach to BC, focusing on the steroid hormone synthesis/metabolism and IGF1 signaling pathways with respect to BC and prostate cancer [28]. Despite the large sample sizes and consequent statistical power gained by the consortium approach used by the BPC3, no highly significant associations have been found between the polymorphisms studied and BC risk [29].

BC GWAS

Since the first GWAS by Klein *et al.* [30], more than 1695 others have been carried out, 27 of which are dedicated to BC [201].

The aim of GWAS is to identify common genetic risk factors by genotyping hundreds of thousands of tagging markers all along the genome. The allelic frequency of each marker is subsequently compared between a group of cases and a group of controls, exactly as is performed in other association studies. The main difference is that GWAS are agnostic: no hypothesis is made as to which variants or genes might be associated with the studied phenotype. Instead, GWAS rely on the knowledge of the structure of the human genome and its regions of linkage disequilibrium (LD). Briefly, LD is a characteristic of markers that tend to co-segregate in a population. Knowing which SNPs are in LD with each other allows the selection of a subset of SNPs, the genotype of which will reflect the variability of the surrounding LD region.

Genome-wide SNP arrays have been developed using the LD information from resources such as the HapMap project [202], maximizing the number of SNPs tagged by the genotyped SNPs. Genome coverage with regard to SNP arrays is therefore expressed as the percentage of known common SNPs (>5%) in strong LD ($r^2 > 0.8$) with at least one SNP on the genotyping array. The coverage has increased with the development of new chips. The Affymetrix SNP Array 5.0, which types approximately 500,000 SNPs, had a global coverage of approximately 65% in European (Centre d'Etude du Polymorphisme Humaine [CEPH] Europeans from Utah) and 41% in African (Yoruban) populations from HapMap [31]. Coverage for Asian populations (Han Chinese + Japanese) is similar to those for European populations. The Illumina Human1M SNP array genotypes more than 1 million SNPs for a coverage of 93 and 68% in the European and the African populations, respectively [31].

The number of individuals included in a GWAS is an extremely important factor, as it will directly impact the statistical power of the analysis. The drawback of agnostic testing of so many markers is the potential for false positives. If one were to assume an association as significant based on the nominal p-value of a test (i.e., $p < 0.01$) while testing 300,000 SNPs, by chance 1% of the tested SNPs would have a p-value lower than this threshold, representing 3000 potentially false-positive associations. It is therefore necessary to use a more stringent threshold ($p < 10^{-7}$ is widely accepted). However, the SNPs that are expected to be identified by GWAS are common variants (minor allele frequency >0.05) with relatively low effect sizes ($1.05 < OR < 2.5$). In order to have enough statistical power to detect such variants, genome-wide studies must include several thousand cases and controls. After a given locus is found associated in a GWAS, further fine mapping must take place to identify the putative functional variant. While the first genome-wide genotyping chips used for BC included approximately 300,000 SNPs, a 1-million SNP chip today costs approximately ten-times less per patient, allowing for cheaper genotyping of larger groups of individuals. Costs associated with recruiting subjects (both in terms of time and money) are still high, however. As a consequence, research groups have developed a variety of GWAS strategies over the years, which are summarized in Table 2.

While GWAS SNPs are probably not the actual functional variant, the nature of the associated loci is of interest. As of 7 September 2013, the GWAS catalog lists 27 BC-associated GWAS [201]. Earlier this year, the Collaborative Oncological Gene–environment Study (COGS) published their findings regarding the identification of 41 novel loci associated with BC risk [32]. This, and related papers, tutorials and other information, is available through the *Nature Genetics* Focus on COGS [203].

Table 2. Strategies for genome-wide association studies.

GWAS strategies	Advantages	Drawbacks	Ref.
Classic multistage	Potential to optimize statistical power and genotyping costs	Needs large numbers of cases and controls	[33,92,93]
Enrich for family history	Increases likelihood that genetic factors are present	Variants can differ between families, and may not be present in the general population	[33,94,95]
Specific tumor type	Reducing heterogeneity in tumor type (among postmenopausal women or restriction to ER^- cancer) can reduce noise	Difficult for rarer tumor types, results not generalizable to the general population	[96–100]
Indirect approach	Using GWAS to identify variants associated with BC risk factors can increase prior probability for certain SNPs	SNPs may be found associated with BC risk factors, but explain little of their contribution to breast cancer incidence	[101–104]

BC: Breast cancer; GWAS: Genome-wide association study; SNP: Single nucleotide polymorphism.

At the time of their identification, the first GWAS loci did not contain genes previously studied in relation with BC, with the exception of *FGFR2* [33], although several had been found to be overexpressed in a variety of cancer cells. Some genes close to GWAS loci are involved in the control of cell growth or cell signaling (e.g., *MAP3K1*), and many have now been linked to diverse cancers through tumor formation, tumor cell dissemination or regulation of apoptosis. GWAS results have therefore been a driving force for the investigation of genes relevant to cancer biology.

The 8q24 'gene desert' is composed of at least five independent loci containing SNPs associated with different cancers [34]. Chromosomal translocations, viral integration and copy number variation at the 8q24 locus have long been observed in many cancers, including BC [35]. More recently, close to 30 SNPs located in this region were found to be associated with risk of various diseases, notably prostate, breast and ovarian cancers, by GWAS [36]. Despite the relative proximity of the *MYC* gene, one of the most studied oncogenes, the biologic mechanisms underlying the identified associations are still unknown, as reports of associations between SNPs and expression levels of *MYC* are contradictory [36]. Adjacent transcribed loci have now been identified (*PTV1*, *PRNCR1*, *miR-1204-1208* and *POU5F1P1*), the investigation of which may bring new information on the functional role(s) of variants associated with breast and other cancers at 8q24 [36].

BC screening

BC screening has been demonstrated to considerably increase diagnosis of breast lesions at an early stage, thus substantially increasing the probability of better prognosis and survival [37–39]. The previously observed increase in incidence of invasive BC in the 1990s has considerably slown down, and between 2001 and 2004 this rate has declined by 3.5% per year [40]. This is mainly attributed to two factors: the increase in BC screening and early detection programs [41]; and the considerable decrease in use of hormone replacement therapy for managing postmenopausal symptoms since 2002 [42,43].

The intensity of screening and medical follow-up depend on a woman's risk estimate as evaluated by their oncologist. To assist them in this task, several tools have been developed [44]. A commonly used tool is the Gail Model [45], which forms the statistical basis of the online tool developed by the National Comprehensive Cancer Network (NCCN) in the USA, and is available online [204]. This model provides an estimation of 5-year and lifetime risk for BC based on the patient's medical and family history of BC, combined with age, ethnicity, age at menarche, age at first full-term pregnancy, number of first-degree relatives diagnosed with BC and whether the patient has already undergone a breast biopsy. Tice *et al.* propose an alternative model that includes mammographic density [46]. Other models, such as Ibis [47] and BODICE [48], are useful for predicting the probability that a woman carries mutations in *BRCA1/2*. Finally, another commonly used tool to detect breast lesions is called the 'triple test', which consists of physical clinical examination, mammography or ultrasound, and fine-needle aspiration. When results of the three steps are concordant, the triple test is estimated to have 100% diagnostic accuracy [49,50].

Based on estimations furnished by this kind of tool, oncologists can decide which type of screening may be best adapted to their patients. Recommended guidelines are not the same for women with low or moderate risk compared with high risk. In the USA, The NCCN, American Cancer Society and US Preventive Services Task Force have published guidelines to help oncologists manage BC screening for patients with moderate to high risk (breast self-examination, clinical breast examination and age to begin mammography screening) [51–53]. The NCCN also recently published an extension to their official BC management guidelines, with specific guidelines for metastatic BC [54]. For women with high risk, more specific guidelines are applied, with earlier screening, more frequent mammographies, or more specific detection techniques such as ultrasound or MRI as a complement to classical mammography screening. However, MRI has disadvantages that might compromise its use as a routine technique, particularly with respect to restrictions on patients who can undergo MRI examination (pacemaker carriers, obese patients and patients with renal failure).

SNPs into clinical practice

One of the earliest promises of GWAS with respect to cancer risk was to improve our ability to identify people in the general population who may benefit from altered screening strategies. As such, it would be possible to detect their cancer sooner, increasing their chances of receiving treatment prior to the tumor spreading. On the other hand, screening methods need

to be very robust, as increasing the number of subjects screened could increase the number of false-positive diagnoses. This would increase burdens on the healthcare system via unnecessary treatments as well as increasing the level of anxiety of patients. Currently, screening is based on age of the patient. In circumstances of strong family history, particularly once mutations in the *BRCA1/2* genes have been detected, screening may be proposed to younger women.

In Britain, it is suggested that women begin organized BC screening at the age of 50 years, and their 10-year risk at this time is 2.3%. Using the seven risk alleles that were known at the time, Pharoah *et al.* calculated the distribution of risk in the general population based on the allele frequencies of the risk alleles [55]. They then used the association between these alleles and BC risk to estimate the age at which the 10-year risk of BC in different percentiles of the population based on this risk was equal to or greater than 2.3%, or that of a 50-year old woman in the general British population. In their example, the women in the top five percentiles of risk would have greater than 2.3% 10-year risk at the age of 41 years, thus potentially making it beneficial to begin screening these women 9 years earlier than recommended.

Despite these relatively positive early findings at the population level, Wacholder *et al.* looked to evaluate the potential for genetic variants to improve on risk models used in the clinic [56]. The Gail Model or Breast Cancer Risk Assessment Tool [45,57] is routinely used to evaluate a woman's BC risk. The Gail Model is based on a woman's reproductive history, family history of BC and previous breast biopsies. In 2010, Wacholder *et al.* added the ten known BC risk variants to the Gail Model. In the data used, the Gail Model plus age, study and entry year had an area under the receiver operating characteristic curve (AUC) of 58%. Including the polymorphisms in the model increased the AUC to only 61.8%. Nearly half of the case subjects (47.2%) were in the same quintile of risk regardless of the model used, while slightly more than 30% of the cases were in a higher quintile and 20% were in a lower quartile. Wacholder *et al.* therefore conclude that these polymorphisms change little with respect to risk prediction [56].

Mealiffe *et al.* used similar methods in a separate data set with very similar results [58]. In this data set, the AUC for the Gail Model went from 56 to 59% after including seven polymorphisms associated with BC risk. While the results of Mealiffe *et al.* [58] are similar to those of Wacholder *et al.* [56], they argue that SNPs may be of interest, particularly for women of intermediate risk as determined by the Gail Model. This conclusion is based on their use of the net reclassification improvement. These analyses revealed that the biggest improvement in risk assessment was among women of intermediate risk. There were also some indications that risk prediction models may vary by estrogen receptor status of BC cases.

More recently, Hüsing *et al.* examined the benefits of using 32 GWAS-identified polymorphisms in risk prediction models [59]. Despite adding a number of additional SNPs, the AUC of the best prediction model including 18 SNPs and epidemiological variables was still only 60.5% and was not statistically different than that observed by Wacholder *et al.* [56]. The group of Pharaoh also updated their analyses, and found that using the GWAS-identified SNPs can reduce the number of women screened [60]. Currently, they estimate that approximately 9% of common BC susceptibility alleles have been discovered.

Using these SNPs could reduce the number of women offered screening based on their 10-year risk by approximately 2%. If all of the BC susceptibility alleles were known, they estimate that a nearly 40% reduction in the number of women offered screening could be achieved [61]. These analyses are no doubt currently being updated with the recent results of the COGS initiative previously mentioned [32].

Future of BC GWAS

After only 5 years of results from GWAS on BC, phenomenal amounts of data have been generated. We have only just begun to exploit this rich resource using classical methods. Over 70 loci associated with BC susceptibility have been identified through GWAS and family studies, which explain approximately 30% of the excess genetic risk of BC, while rare and high-to-moderate-risk loci contribute approximately 20% [32]. A great deal of work still needs to be carried out to fully exploit these data. Many more loci are suspected to contribute to the polygenic susceptibility to BC, most of which are likely tagged by existing GWAS chips, but their marginal significance does not reach the required genome-wide significance threshold of $p < 10^{-7}$. In other words, a number of false-negative associations have likely been excluded based on current GWAS data. In order to extract more biologically relevant

information from GWAS, analysis methods able to sift true associations from these false negatives are required. The most promising avenues to follow with respect to a more complete view of genetic susceptibility to BC are:

- Gene–environment (G×E) interactions;
- Gene–gene (G×G) interactions;
- Alternative statistical analyses.

As shown in TABLE 1, a number of nongenetic factors are associated with BC risk. It has been proposed that leveraging nongenetic risks and G×E interactions can add power to GWAS scans [62]. Given that the majority of BC GWAS will have a large quantity of environmental and lifestyle factors already available for the majority of their participants, carrying out G×E interaction studies in this context requires only additional computer and analyst time. This approach has been used by a number of studies [63,64]. However, the statistical limitations of G×E studies in terms of GWAS are daunting. Examining the interactions between even one nongenetic exposure and the SNPs of a GWAS at least doubles the number of statistical tests carried out, reducing statistical power and the capacity to detect associations. However, a number of groups have developed methods to overcome the statistical limitations of the increased number of tests carried out in interaction analyses [65,66].

Given the relatively high incidence of BC and the paucity of polymorphisms found to be more moderately associated (OR >1.5) with the disease, it has been proposed that combinations of variants, or G×G interactions, are necessary to explain the genetic component of BC risk. In more classical genetic terms, this phenomenon is known as epistasis. Similar to G×E interactions, examining G×G in the GWAS context would exponentially increase the number of statistical tests to be carried out, and, as such, drastically reduce power. Despite this drawback, statistical methods to evaluate G×G in GWAS have been developed, including but not limited to BOOST [67], permutation testing [68] and multifactor dimension reduction [69].

A variety of approaches to further statistically exploit GWAS data already exist. In line with the agnostic nature of GWAS, LASSO-like methods use multivariate regressions together with variable reduction and selection to let the genotype data lead to the most influential SNPs [70,71]. The output of purely mathematical approaches can be extremely difficult to interpret and replicate. Other LASSO-like methods that also integrate publicly available biological information such as biological pathways exist [72]. Alternatively, a wide range of methods are dedicated to the inclusion of biological knowledge based on enrichment of association signal in biological pathways [73,74] or within modules in protein–protein interaction networks [75]. Specifically, Braun and Buetow used their pathway analyses to confirm the association between *FGFR2* and BC risk, as well as novel putative associations between purine metabolism, ERbB signaling, calcium signaling and GnRH signaling pathways, and BC risk [76]. Finally, Bayesian statistics allow the integration of biological knowledge directly within the statistical modeling of the data [76]. Using the framework developed by Wakefield, one can integrate biological information of any kind and produce a reordering of the GWAS SNPs according to both their genotypes and prior knowledge to discover real associations that would have been missed by classical analysis [77].

Integration of other sources of biological information such as regulatory annotations from genome-wide chromatin immunoprecipitation sequencing or footprinting methods generated by international consortia such as ENCODE will certainly prove very useful [78]. Recently, a study combining BC cell transcription factor binding sites for *FOX1A* and *ERα* with BC GWAS hits noted an enrichment in associated loci within the binding sites, as well as an effect of the majority of the risk-associated SNPs on the affinity of FOXA1 for chromatin [79]. Li *et al.* examined data from the TCGA and identified a number of variants that act through influencing the expression of *ESR1*, *MYC* and *KLF4* [80]. These studies prove the functional importance of GWAS-identified loci for BC and the potential of combining GWAS data together with regulatory annotations obtained through modern, cell-type specific OMIC techniques.

Conclusion

There is no question that the field of genomics is rapidly advancing and evolving. The costs of genotyping and, more importantly, next-generation sequencing, continue to drop. A number of large-scale whole-genome sequencing projects have been initiated to describe the global variation present in different tumor types, such as the TCGA [205] and ICGC [206]. Concurrently, next-generation sequencing has entered the clinic, with trials aimed at

the evaluation of targeted sequencing of genes and mutational hotspots as a means to provide personalized therapy for cancer patients currently underway. However, with this increase in capacity to generate genetic data comes an increase in the need for tools and resources to analyze the deluge of data generated by these studies. As discussed in the previous section, a great deal of information remains to be gleaned from existing GWAS. While sequencing may allow the detection of very rare variants that may influence disease risk, the sheer volume of work needed to identify and validate these variants limits our ability to rapidly exploit the data. Furthermore, classical association testing is not adapted to evaluate the influence of rare variants on disease. It is therefore important to continue to invest resources to fully exploit existing GWAS data, as well as to apply this still powerful technology to understudied aspects of cancer epidemiology.

Future perspective

GWAS is still a relatively young technology that is only starting to mature. As further developments in the identification of variants associated with human phenotypes are made, transferring these results to clinical applications will become possible. We are therefore likely on the verge of realizing at least some of the promise initially anticipated from GWAS studies with respect to personalizing medicine.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

Executive summary

Epidemiology of breast cancer

- Breast cancer is the most common cancer among women in developed countries, and is a leading cause of cancer death among women. Incidence rates are rising, and breast cancer will become a major public health concern in developing countries as improvements in infant's and women's health issues lead to increased life expectancy. A number of risk factors have been identified as being associated with breast cancer, including genetic variants.

Genetic variants & breast cancer risk

- A positive family history of breast cancer increases one's personal risk of developing the disease. However, a small portion of breast cancer cases are due to moderate- or high-penetrance mutations. Early work on genetic variants and breast cancer risk has led to only a few positive associations, despite having strong biology-based hypotheses.

Breast cancer genome-wide association studies

- The explosion of our potential to explore the human genome based largely on the first full genome sequences being published in the early 2000s has led to the development of the genome-wide association studies (GWAS). Nearly 2000 GWAS have been published, with 27 of these dedicated to breast cancer. These studies are based on linkage disequilibrium, or correlation, between genetic variants. They use an agnostic approach to screen the genome for variants associated with measurable phenotypes, including cancer risk. A number of alternative approaches for GWAS design have been proposed. Given their size and cost, a number of considerations for GWAS need to be taken into consideration. The first breast cancer GWAS were published in 2007, and since then a number of additional studies have been carried out.

Breast cancer screening

- Currently, breast cancer screening is offered based solely on a woman's age. However, a number of models have been developed to measure breast cancer risk. These models take into consideration a number of factors, including family history, reproductive history and breast density. However, these models are rarely used in a clinical setting.

Applying single nucleotide polymorphisms to clinical practice in screening

- One of the greatest promises of GWAS was to identify people who may benefit from altered screening strategies based on their risk. A number of studies have evaluated the potential for single nucleotide polymorphisms identified through breast cancer GWAS to better measure breast cancer risk. These models were built based on earlier GWAS results, and the most recent wave of results has yet to be included.

Future of breast cancer GWAS

- Despite the exponential growth in our capacity to analyze genomes, the wealth of data generated from GWAS studies remains to be fully exploited. Examination of gene–gene and gene–environment interactions needs to be explored. Alternative statistical analyses may be carried out in order to identify further novel polymorphisms associated with breast cancer risk. Finally, the integration of additional biological information into the analyses of GWAS data may lead to further insights into the genetic component of breast cancer risk.

References

Papers of special note have been highlighted as:
▪ of interest
▪▪ of considerable interest

1 Curado MP. Breast cancer in the world: incidence and mortality. *Salud Publica Mex.* 53(5), 372–384 (2011).

2 Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int. J. Cancer* 127(12), 2893–2917 (2010).

3 Writing Group for the American Cancer Society. *Breast Cancer, Facts & Figures 2011–2012.* American Cancer Society, GA, USA (2012).

4 Writing Group for the American Cancer Society. *Cancer Facts & Figures for African Americans 2011–2012.* American Cancer Society, GA, USA (2012).

5 Writing Group for the American Cancer Society. *Cancer Facts & Figures for Hispanics/ Latinos 2009–2011.* American Cancer Society, GA, USA (2011).

6 Song C, Chen GK, Millikan RC *et al.* A genome-wide scan for breast cancer risk haplotypes among African American women. *PLoS ONE* 8(2), e57298 (2013).

7 Rinella ES, Shao Y, Yackowski L *et al.* Genetic variants associated with breast cancer risk for Ashkenazi Jewish women with strong family histories but no identifiable *BRCA1/2* mutation. *Hum. Genet.* 132(5), 523–536 (2013).

8 Ziegler RG, Hoover RN, Pike MC *et al.* Migration patterns and breast cancer risk in Asian–American women. *J. Natl Cancer Inst.* 85(22), 1819–1827 (1993).

9 Lalloo F, Evans DG. Familial breast cancer. *Clin. Genet.* 82(2), 105–114 (2012).

10 Newman B, Austin MA, Lee M, King MC. Inheritance of human breast cancer: evidence for autosomal dominant transmission in high-risk families. *Proc. Natl Acad. Sci. USA* 85(9), 3044–3048 (1988).

11 Hall JM, Lee MK, Newman B *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* 250(4988), 1684–1689 (1990).

12 Miki Y, Swensen J, Shattuck-Eidens D *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1. Science* 266(5182), 66–71 (1994).

13 Kurian AW. *BRCA1* and *BRCA2* mutations across race and ethnicity: distribution and clinical implications. *Curr. Opin. Obstet. Gynecol.* 22(1), 72–78 (2010).

14 Antoniou AC, Cunningham AP, Peto J *et al.* The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. *Br. J. Cancer* 98(8), 1457–1466 (2008).

15 Evans DG, Shenton A, Woodward E, Lalloo F, Howell A, Maher ER. Penetrance estimates for *BRCA1* and *BRCA2* based on genetic testing in a clinical cancer genetics service setting: risks of breast/ovarian cancer quoted should reflect the cancer burden in the family. *BMC Cancer* 8, 155 (2008).

16 Antoniou A, Pharoah PD, Narod S *et al.* Average risks of breast and ovarian cancer associated with *BRCA1* or *BRCA2* mutations detected in case series unselected for family history: a combined analysis of 22 studies. *Am. J. Hum. Genet.* 72(5), 1117–1130 (2003).

17 Writing Group for the Anglian Breast Cancer Study Group. Prevalence and penetrance of *BRCA1* and *BRCA2* mutations in a population-based series of breast cancer cases. Anglian Breast Cancer Study Group. *Br. J. Cancer* 83(10), 1301–1308 (2000).

18 Tan MH, Mester JL, Ngeow J, Rybicki LA, Orloff MS, Eng C. Lifetime cancer risks in individuals with germline *PTEN* mutations. *Clin. Cancer Res.* 18(2), 400–407 (2012).

19 Hearle N, Schumacher V, Menko FH *et al.* Frequency and spectrum of cancers in the Peutz–Jeghers syndrome. *Clin. Cancer Res.* 12(10), 3209–3215 (2006).

20 Pharoah PD, Guilford P, Caldas C; International Gastric Cancer Linkage Consortium. Incidence of gastric cancer and breast cancer in *CDH1* (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. *Gastroenterology* 121(6), 1348–1353 (2001).

21 Ripperger T, Gadzicki D, Meindl A, Schlegelberger B. Breast cancer susceptibility: current knowledge and implications for genetic counselling. *Eur. J. Hum. Genet.* 17(6), 722–731 (2008).

22 Thompson D, Duedal S, Kirner J *et al.* Cancer risks and mortality in heterozygous ATM mutation carriers. *J. Natl Cancer Inst.* 97(11), 813–822 (2005).

23 Meijers-Heijboer H, van den Ouweland A, Klijn J *et al.* Low-penetrance susceptibility to breast cancer due to *CHEK2*(*)1100delC in noncarriers of *BRCA1* or *BRCA2* mutations. *Nat. Genet.* 31(1), 55–59 (2002).

24 Seal S, Thompson D, Renwick A *et al.* Truncating mutations in the Fanconi anemia J gene *BRIP1* are low-penetrance breast cancer susceptibility alleles. *Nat. Genet.* 38(11), 1239–1241 (2006).

25 Rahman N, Seal S, Thompson D *et al. PALB2*, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat. Genet.* 39(2), 165–167 (2007).

26 Cox A, Dunning AM, Garcia-Closas M *et al.* A common coding variant in *CASP8* is associated with breast cancer risk. *Nat. Genet.* 39(3), 352–358 (2007).

27 Cox DG, Penney K, Guo Q, Hankinson SE, Hunter DJ. *TGFB1* and *TGFBR1* polymorphisms and breast cancer risk in the nurses' health study. *BMC Cancer* 7, 175 (2007).

28 Hunter DJ, Riboli E, Haiman CA *et al.* A candidate gene approach to searching for low-penetrance breast and prostate cancer genes. *Nat. Rev. Cancer* 5(12), 977–985 (2005).

29 Canzian F, Cox DG, Setiawan VW *et al.* Comprehensive analysis of common genetic variation in 61 genes related to steroid hormone and insulin-like growth factor-I metabolism and breast cancer risk in the NCI breast and prostate cancer cohort consortium. *Hum. Mol. Genet.* 19(19), 3873–3884 (2010).

30 Klein RJ, Zeiss C, Chew EY *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* 308(5720), 385–389 (2005).

31 Li M, Li C, Guan W. Evaluation of coverage variation of SNP chips for genome-wide association studies. *Eur. J. Hum. Genet.* 16(5), 635–643 (2008).

32 Michailidou K, Hall P, Gonzalez-Neira A *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* 45(4), 353–361 (2013).

▪▪ **Presents significant results for a large number of single nucleotide polymorphisms associated with breast cancer risk.**

33 Easton DF, Pooley KA, Dunning AM *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 447(7148), 1087–1093 (2007).

34 Ghoussaini M, Song H, Koessler T *et al.* Multiple loci with different cancer specificities within the 8q24 gene desert. *J. Natl Cancer Inst.* 100(13), 962–966 (2008).

35 Yokota T, Yoshimoto M, Akiyama F *et al.* Frequent multiplication of chromosomal region 8q24.1 associated with aggressive histologic types of breast cancers. *Cancer Lett.* 139(1), 7–13 (1999).

36 Huppi K, Pitt JJ, Wahlberg BM, Caplen NJ. The 8q24 gene desert: an oasis of non-coding transcriptional activity. *Front. Genet.* 3, 69 (2012).

37 Moss SM, Cuckle H, Evans A, Johns L, Waller M, Bobrow L; Trial Management Group. Effect of mammographic screening from age 40 years on breast cancer mortality at 10 years' follow-up: a randomised controlled trial. *Lancet* 368(9552), 2053–2060 (2006).

38 Smith RA, Duffy SW, Gabe R, Tabar L, Yen AM, Chen TH. The randomized trials of breast cancer screening: what have we learned? *Radiol. Clin. North Am.* 42(5), 793–806, v (2004).

39 Tabár L, Vitak B, Chen TH *et al.* Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. *Radiology* 260(3), 658–663 (2011).

40 Writing Group for the American Cancer Society. *Breast Cancer, Facts & Figures 2007–2008*. American Cancer Society, GA, USA (2008).

41 Bleyer A, Welch HG. Effect of three decades of screening mammography on breast-cancer incidence. *N. Engl. J. Med.* 367(21), 1998–2012 (2005).

42 Chen FP. Postmenopausal hormone therapy and risk of breast cancer. *Chang Gung Med. J.* 32(2), 140–147 (2009).

43 Gompel A, Santen RJ. Hormone therapy and breast cancer risk 10 years after the WHI. *Climacteric* 15(3), 241–249 (2012).

44 Rim A, Chellman-Jeffers M. Trends in breast cancer screening and diagnosis. *Cleve. Clin. J. Med.* 75(Suppl. 1), S2–S9 (2008).

45 Gail MH, Brinton LA, Byar DP *et al.* Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl Cancer Inst.* 81(24), 1879–1886 (1989).

46 Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerlikowske K. Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. *Ann. Intern. Med.* 148(5), 337–347 (2008).

47 Cuzick J, Forbes J, Edwards R *et al.* First results from the international breast cancer intervention study (IBIS-I): a randomised prevention trial. *Lancet* 360(9336), 817–824 (2002).

48 Antoniou AC, Hardy R, Walker L *et al.* Predicting the likelihood of carrying a *BRCA1* or *BRCA2* mutation: validation of BOADICEA, BRCAPRO, IBIS, Myriad and the Manchester scoring system using data from UK genetics clinics. *J. Med. Genet.* 45(7), 425–431 (2008).

49 Vetto J, Pommier R, Schmidt W *et al.* Use of the "triple test" for palpable breast lesions yields high diagnostic accuracy and cost savings. *Am. J. Surg.* 169(5), 519–522 (1995).

50 Vetto JT, Pommier RF, Schmidt WA, Eppich H, Alexander PW. Diagnosis of palpable breast lesions in younger women by the modified triple test is accurate and cost-effective. *Arch. Surg.* 131(9), 967–972; discussion 972–974 (1996).

51 Smith RA, Saslow D, Sawyer KA *et al.* American cancer society guidelines for breast cancer screening: update 2003. *CA Cancer J. Clin.* 53(3), 141–169 (2003).

52 Bevers TB, Anderson BO, Bonaccio E *et al.* NCCN clinical practice guidelines in oncology: breast cancer screening and diagnosis. *J. Natl Compr. Canc. Netw.* 7(10), 1060–1096 (2009).

53 Writing Group for the U.S. Preventive Services Task Force. Screening for breast cancer: recommendations and rationale. *Ann. Intern. Med.* 137(5 Pt 1), 344–346 (2002).

54 Carlson RW, Allred DC, Anderson BO *et al.* Metastatic breast cancer, version 1.2012: featured updates to the NCCN guidelines. *J. Natl Compr. Canc. Netw.* 10(7), 821–829 (2012).

55 Pharoah PD, Antoniou AC, Easton DF, Ponder BA. Polygenes, risk prediction, and targeted prevention of breast cancer. *N. Engl. J. Med.* 358, 2796–2803 (2008).

▪ **Forms the groundwork for evaluating the use of single nucleotide polymorphisms in breast cancer screening, and will probably be updated, given recent findings with respect to polymorphisms associated with breast cancer risk.**

56 Wacholder S, Hartge P, Prentice R *et al.* Performance of common genetic variants in breast-cancer risk models. *N. Engl. J. Med.* 362(11), 986–993 (2010).

57 Gail MH, Benichou J. Validation studies on a model for breast cancer risk. *J. Natl Cancer Inst.* 86(8), 573–575 (1994).

58 Mealiffe ME, Stokowski RP, Rhees BK, Prentice RL, Pettinger M, Hinds DA. Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. *J. Natl Cancer Inst.* 102(21), 1618–1627 (2010).

59 Hüsing A, Canzian F, Beckmann L *et al.* Prediction of breast cancer risk by genetic risk factors, overall and by hormone receptor status. *J. Med. Genet.* 49(9), 601–608 (2012).

60 Gurley WB, Kemp JD, Albert MJ, Sutton DW, Callis J. Transcription of Ti plasmid-derived sequences in three octopine-type crown gall tumor lines. *Proc. Natl. Acad. Sci. USA* 76(6), 2828–2832 (1979)

61 Ghoussaini M, Fletcher O, Michailidou K *et al.* Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nat. Genet.* 44(3), 312–318 (2012).

62 Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene–environment interaction to detect genetic associations. *Hum. Hered.* 63(2), 111–119 (2007).

63 Milne RL, Gaudet MM, Spurdle AB *et al.* Assessing interactions between the associations of common genetic susceptibility variants, reproductive history and body mass index with breast cancer risk in the breast cancer association consortium: a combined case–control study. *Breast Cancer Res.* 12(6), R110 (2010).

64 Nickels S, Truong T, Hein R *et al.* Evidence of gene–environment interactions between common breast cancer susceptibility loci and established environmental risk factors. *PLoS Genet.* 9(3), e1003284 (2013).

65 Hein R, Beckmann L, Chang-Claude J. Sample size requirements for indirect association studies of gene–environment interactions (G × E). *Genet. Epidemiol.* 32(3), 235–245 (2008).

66 Murcray CE, Lewinger JP, Conti DV, Thomas DC, Gauderman WJ. Sample size requirements to detect gene–environment interactions in genome-wide association studies. *Genet. Epidemiol.* 35(3), 201–210 (2011).

67 Wan X, Yang C, Yang Q *et al.* BOOST: a fast approach to detecting gene–gene interactions in genome-wide case–control studies. *Am. J. Hum. Genet.* 87(3), 325–340 (2010).

68 Wongseree W, Assawamakin A, Piroonratana T, Sinsomros S, Limwongse C, Chaiyaratana N. Detecting purely epistatic multi-locus interactions by an omnibus permutation test on ensembles of two-locus analyses. *BMC Bioinformatics* 10, 294 (2009).

69 Oh S, Lee J, Kwon MS, Weir B, Ha K, Park T. A novel method to identify high order gene–gene interactions in genome-wide association studies: gene-based MDR. *BMC Bioinformatics* 13(Suppl. 9), S5 (2012).

70 Tibshirani R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* 58(1), 267–288 (1996).

71 Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 25(6), 714–721 (2009).

72 Chen LS, Hutter CM, Potter JD *et al.* Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *Am. J. Hum. Genet.* 86(6), 860–871 (2010).

73 Menashe I, Maeder D, Garcia-Closas M *et al.* Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. *Cancer Res.* 70(11), 4453–4459 (2010).

74 Braun R, Buetow K. Pathways of distinction analysis: a new technique for multi-SNP analysis of GWAS data. *PLoS Genet.* 7(6), e1002101 (2011).

75 Jia P, Zheng S, Long J, Zheng W, Zhao Z. dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks. *Bioinformatics* 27(1), 95–102 (2011).

76 Wakefield J. A bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* 81(2), 208–227 (2007).

77 Johansson M, Roberts A, Chen D *et al.* Using prior information from the medical literature in GWAS of oral cancer identifies novel susceptibility variant on chromosome 4 – the AdAPT method. *PLoS ONE* 7(5), e36888 (2012).

78 Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. *Genome Res.* 22(9), 1748–1759 (2012).

79 Cowper-Sal lari R, Zhang X, Wright JB *et al.* Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.* 44(11), 1191–1198 (2012).

80 Qiyuan Li, Ji-Heui Seo, Barbara Stranger *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* 152(3), 633–641 (2013).

▪ **Demonstrates how expression analyses can be used in the genome-wide association study context.**

81 Cox DG, Dostal L, Hunter DJ *et al.* *N*-acetyltransferase 2 polymorphisms, tobacco smoking, and breast cancer risk in the breast and prostate cancer cohort consortium. *Am. J. Epidemiol.* 174(11), 1316–1322 (2011).

82 Coronado GD, Beasley J, Livaudais J. Alcohol consumption and the risk of breast cancer. *Salud Publica Mex.* 53(5), 440–447 (2011).

83 Seitz HK, Pelucchi C, Bagnardi V, La Vecchia C. Epidemiology and pathophysiology of alcohol and breast cancer: update 2012. *Alcohol Alcohol.* 47(3), 204–212 (2012).

84 Rossouw JE, Anderson GL, Prentice RL *et al.*; Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. *JAMA* 288(3), 321–333 (2002).

85 Vessey M, Painter R. Oral contraceptive use and cancer. findings in a large cohort study, 1968–2004. *Br. J. Cancer* 95(3), 385–389 (2006).

86 Hannaford PC, Selvaraj S, Elliott AM, Angus V, Iversen L, Lee AJ. Cancer risk among users of oral contraceptives: cohort data from the royal college of general practitioner's oral contraception study. *BMJ* 335(7621), 651 (2007).

87 Cibula D, Gompel A, Mueck AO *et al.* Hormonal contraception and risk of cancer. *Hum. Reprod. Update* 16(6), 631–650 (2010).

88 Kahlenborn C, Modugno F, Potter DM, Severs WB. Oral contraceptive use as a risk factor for premenopausal breast cancer: a meta-analysis. *Mayo Clin. Proc.* 81(10), 1290–1302 (2006).

89 Renehan AG, Tyson M, Egger M, Heller RF, Zwahlen M. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *Lancet* 371(9612), 569–578 (2008).

90 McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. *Cancer Epidemiol. Biomarkers Prev.* 15(6), 1159–1169 (2006).

91 Dumas I, Diorio C. Estrogen pathway polymorphisms and mammographic density. *Anticancer Res.* 31(12), 4369–4386 (2011).

92 Hunter DJ, Kraft P, Jacobs KB. Jacobs *et al.* A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* 39(7), 870–874 (2007).

93 Stacey SN, Manolescu A, Sulem P *et al.* Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.* 39(7), 865–869 (2007).

94 Turnbull C, Ahmed S, Morrison J *et al.* Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.* 42(6), 504–507 (2010).

95 Fletcher O, Johnson N, Orr N *et al.* Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. *J. Natl Cancer Inst.* 103(5), 425–435 (2011).

96 Li J, Humphreys K, Darabi H *et al.* A genome-wide association scan on estrogen receptor-negative breast cancer. *Breast Cancer Res.* 12(6), R93 (2010).

97 Haiman CA, Chen GK, Vachon CM *et al.* A common variant at the *TERT–CLPTM1L* locus is associated with estrogen receptor-negative breast cancer. *Nat. Genet.* 43(12), 1210–1214 (2011).

98 Siddiq A, Couch FJ, Chen GK *et al.* A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. *Hum. Mol. Genet.* 21(24), 5373–5384 (2012).

99 Stevens KN, Vachon CM, Lee AM *et al.* Common breast cancer susceptibility loci are associated with triple-negative breast cancer. *Cancer Res.* 71(19), 6240–6249 (2011).

100 Elgazzar S, Zembutsu H, Takahashi A *et al.* A genome-wide association study identifies a genetic variant in the *SIAH2* locus associated with hormonal receptor-positive breast cancer in Japanese. *J. Hum. Genet.* 57(12), 766–771 (2012).

101 Kaplan RC, Petersen AK, Chen MH *et al.* A genome-wide association study identifies novel loci associated with circulating IGF-I and IGFBP-3. *Hum. Mol. Genet.* 20(6), 1241–1251 (2011).

102 Prescott J, Thompson DJ, Kraft P *et al.* Genome-wide association study of circulating estradiol, testosterone, and sex hormone-binding globulin in postmenopausal women. *PLoS ONE* 7(6), e37815 (2012).

103 Stevens KN, Lindstrom S, Scott CG *et al.* Identification of a novel percent mammographic density locus at 12q24. *Hum. Mol. Genet.* 21(14), 3299–3305 (2012).

104 Varghese JS, Thompson DJ, Michailidou K *et al.* Mammographic breast density and breast cancer: evidence of a shared genetic basis. *Cancer Res.* 72(6), 1478–1484 (2012).

Websites

201 National Human Genome Research Institute (NHGRI). Division of Genomic Medicine. A Catalog of Published Genome-Wide Association Studies. www.genome.gov/gwastudies

202 International HapMap Project. http://hapmap.ncbi.nlm.nih.gov

203 Nature. iCOGS. www.nature.com/icogs

204 National Cancer Institute (NCI). Breast Cancer Risk Assessment Tool. http://cancer.gov/bcrisktool

205 The Cancer Genome Atlas. http://cancergenome.nih.gov

206 The International Cancer Genome Consortium (ICGC). http://icgc.org

An original phylogenetic approach identified mitochondrial haplogroup T1a1 as inversely associated with breast cancer risk in *BRCA2* mutation carriers.

Sophie Blein[1], Claire Bardel[2,3,4], Vincent Danjean[5,6], Lesley McGuffog[7], Sue Healey[8], Daniel Barrowdale[7], Andrew Lee[7], Joe Dennis[7], Karoline B. Kuchenbaecker[7], Penny Soucy[9], Mary Beth Terry[10], Wendy K. Chung[11], David E. Goldgar[12], Saundra S. Buys[13], BCFR[14], Ramunas Janavicius[15], Laima Tihomirova[16], Nadine Tung[17], Cecilia M. Dorfling[18], Elizabeth J. van Rensburg[18], Susan L. Neuhausen[19], Yuan Chun Ding[19], Anne-Marie Gerdes[20], Bent Ejlertsen[21], Finn C. Nielsen[22], Thomas v. O. Hansen[22], Ana Osorio[23], Javier Benitez[24], Raquel Andrés-Conejero[25], Ena Segota[26], Jeffrey N. Weitzel[27], Margo Thelander[28], Paolo Peterlongo[29], Paolo Radice[30], Valeria Pensotti[31], Riccardo Dolcetti[32], Bernardo Bonanni[33], Bernard Peissel[34], Daniela Zaffaroni [34], Giulietta Scuvera[34], Siranoush Manoukian[34], Liliana Varesco[35], Gabriele L Capone[36], Laura Papi[37], Laura Ottini[38], Drakoulis Yannoukakos[39], Irene Konstantopoulou[40], Judy Garber[41], Ute Hamann[42], Alan Donaldson[43], Angela Brady [44], Carole Brewer[45], Claire Foo[46], D. Gareth Evans[47], Debra Frost[48], Diana Eccles[49], EMBRACE[48], Fiona Douglas[50], Jackie Cook[51], Julian Adlard[52], Julian Barwell[53], Lisa Walker[54], Louise Izatt[55], Lucy E. Side [56], M. John Kennedy[57], Marc Tischkowitz[58], Mark T. Rogers[59], Mary E. Porteous[60], Patrick J. Morrison[61], Radka Platte[48], Ros Eeles[62], Rosemarie Davidson[63], Shirley Hodgson[64], Trevor Cole[65], Andrew K. Godwin[66], Claudine Isaacs[67], Kathleen Claes[68], Kim De Leeneer[68], Alfons Meindl [69], Andrea Gehrig[70], Barbara Wappenschmidt[71], Christian Sutter[72], Christoph Engel[73], Dieter Niederacher[74], Doris Steinemann[75], Hansjoerg Plendl[76], Karin Kast[77], Kerstin Rhiem[71], Nina Ditsch[69], Norbert Arnold[78], Raymonda Varon-Mateeva[79], Rita K. Schmutzler[80], Sabine Preisler-Adams[81], Shan Wang-Gohrke[82], Antoine de Pauw[83], Cédrick Lefol[83], Christine Lasset[84, 85], Dominique Leroux[86, 87], Etienne Rouleau[88], Francesca Damiola[1], GEMO Study Collaborators[89], Hélène Dreyfus[87,90], Laure Barjhoux[1], Lisa Golmard [83], Nancy Uhrhammer[91], Valérie Bonadona[85,92], Valérie Sornin[1], Yves-Jean Bignon[91], Jonathan Carter[93], Linda Van Le[94], Marion Piedmonte[95], Paul A. DiSilvestro[96], Miguel de la Hoya[97], Trinidad Caldes[97], Heli Nevanlinna[98], Kristiina Aittomäki[99], Agnes Jager[100], Ans M.W. van den Ouweland[101], Carolien M. Kets[102], Cora M. Aalfs[103], Flora E. van Leeuwen[104], Frans B.L. Hogervorst[105], Hanne E.J. Meijers-Heijboer[106], HEBON[107], Jan C. Oosterwijk[108], Kees E.P. van Roozendaal[109], Matti A. Rookus[104], Peter Devilee[110], Rob B. van der Luijt[111], Edith Olah[112], Orland Diez [113], Alex Teulé[114], Conxi Lazaro[115], Ignacio Blanco[114], Jesús Del Valle[115], Anna Jakubowska[116], Grzegorz Sukiennicki[117], Jacek Gronwald[116], Jan Lubinski[116], Katarzyna Durda [116], Katarzyna Jaworska-Bieniek[117], Bjarni A. Agnarsson[118], Christine Maugard[119], Alberto Amadori[120], Marco Montagna[121], Manuel R. Teixeira[122], Amanda B. Spurdle[8], William Foulkes[123], Curtis Olswold[124], Noralane Lindor[125], Vernon S. Pankratz[124], Csilla I. Szabo[126], Anne Lincoln[127], Lauren Jacobs[127], Marina Corines[127], Mark Robson[128], Joseph Vijai[128], Andreas Berger[129], Anneliese Fink-Retter[129], Christian F. Singer[129], Christine Rappaport[129], Daphne Geschwantler Kaulich[129], Georg Pfeiler[130], Muy-Kheng Tea[129], Mark H. Greene[131], Phuong L. Mai[132], Gad Rennert1[33], Evgeny N. Imyanitov[134], Anna Marie Mulligan[135], Gord Glendon[136], Irene L. Andrulis[137], Sandrine Tchatchou[138], Amanda Ewart Toland[139], Inge Sokilde Pedersen[140], Mads Thomassen[141], Torben A. Kruse[141], Uffe Birk Jensen[142], Maria A. Caligo[143], Eitan Friedman[144], Jamal Zidan [145], Yael Laitman [144], Annika Lindblom[146], Beatrice Melin[147], Brita Arver[148], Niklas Loman[149], Richard Rosenquist[150], Olufunmilayo I. Olopade[151], Robert L. Nussbaum[152], Susan J. Ramus[153], Katherine L. Nathanson[154], Susan M. Domchek[154], Timothy R. Rebbeck[155], Banu K. Arun[156], Gillian Mitchell[157], Beth Y. Karlan[158], Jenny Lester[158], Sandra Orsulic[158], Dominique Stoppa-Lyonnet[159,160,161], Gilles Thomas[162], Jacques Simard[9], Fergus J. Couch[163], Kenneth Offit[128], Douglas F. Easton[7], Georgia Chenevix-Trench[8], Antonis C. Antoniou[7], Sylvie Mazoyer[1], Catherine M. Phelan[164], Olga M. Sinilnikova[1,165], David G. Cox[1]

Affiliations :

[1] INSERM U1052, CNRS UMR5286, Université Lyon 1, Centre de Recherche en Cancérologie de Lyon, Lyon, France

[2] Université de Lyon, F-69000 Lyon, France

[3] Université Lyon 1, F-69100, Villeurbanne, France
[4] CNRS UMR5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique-Santé, F-69100, Villeurbanne, France
[5] Univ. Grenoble Alpes, UMR 5217, Laboratoire LIG, équipe MOAIS, F-38041 Grenoble, France
[6] INRIA Rhône-Alpes, équipe-projet MOAIS, F-38334 Saint Ismier Cedex, France
[7] Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
[8] Department of Genetics and Computational Biology, QIMR Berghofer, Brisbane, Australia
[9] Centre Hospitalier Universitaire de Québec Research Center and Laval University, Quebec City, Canada
[10] Department of Epidemiology, Columbia University, New York, NY, USA
[11] Departments of Pediatrics and Medicine, Columbia University, New York, NY, USA
[12] Department of Dermatology, University of Utah School of Medicine, Salt Lake City, Utah, USA
[13] Department of Internal Medecine, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, USA
[14] Department of Epidemiology, Cancer Prevention Institute of California, Fremont, California
[15] Vilnius University Hospital Santariskiu Clinics, Hematology, oncology and transfusion medicine center, Dept. of Molecular and Regenerative Medicine; State Research Institute Centre for Innovative medicine, Vilnius, Lithuania
[16] Latvian Biomedical Research and Study Centre, Riga, Latvia
[17] Department of Medical Oncology, Beth Israel Deaconess Medical Center
[18] Department of Genetics, University of Pretoria, Pretoria, South Africa
[19] Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, CA USA
[20] Department of Clincial Genetics, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark
[21] Department of Oncology, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark
[22] Center for Genomic Medicine, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark
[23] Human Genetics Group, Spanish National Cancer Centre (CNIO), Madrid, Spain, and Biomedical Network on Rare Diseases (CIBERER).
[24] Human Genetics Group and Genotyping Unit, Spanish National Cancer Centre (CNIO), Madrid, Spain, and Biomedical Network on Rare Diseases (CIBERER).
[25] Medical Oncology Service, Hospital Clínico Lozano Blesa. San Juan Bosco 15 50009. Zaragoza, Spain
[26] Holy Cross Hospital-Michael and Dianne Bienes Cancer Center, care of City of Hope Clinical Cancer Genetics Community Research Network
[27] Clinical Cancer Genetics, City of Hope, 1500 East Duarte Road, Duarte, California 91010 USA (for the City of Hope Clinical Cancer Genetics Community Research Network)
[28] John Muir Medical Center, care of City of Hope Clinical Cancer Genetics Community Research Network
[29] IFOM, Fondazione Istituto FIRC di Oncologia Molecolare, Milan, Italy.
[30] Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Preventive and Predictive Medicine, Fondazione IRCCS Istituto Nazionale Tumori (INT), Milan, Italy
[31] IFOM, Fondazione Istituto FIRC di Oncologia Molecolare and Cogentech Cancer Genetic Test Laboratory, Milan, Italy
[32] Cancer Bioimmunotherapy Unit, CRO Aviano National Cancer Institute, Aviano (PN), Italy
[33] Division of Cancer Prevention and Genetics, Istituto Europeo di Oncologia, Milan, Italy
[34] Unit of Medical Genetics, Department of Preventive and Predictive Medicine, Fondazione IRCCS Istituto Nazionale Tumori (INT), Milan, Italy
[35] Unit of Hereditary Cancer, Department of Epidemiology, Prevention and Special Functions, IRCCS AOU San Martino - IST Istituto Nazionale per la Ricerca sul Cancro, Genoa, Italy
[36] Unit of Medical Genetics, Department of Biomedical, Experimental and Clinical Sciences, University of Florence, Florence, Italy and FiorGen Foundation for Pharmacogenomics, Sesto Fiorentino (FI), Italy

[37] Unit of Medical Genetics, Department of Biomedical, Experimental and Clinical Sciences, University of Florence, Florence, Italy
[38] Department of Molecular Medicine, Sapienza University, Rome, Italy
[39] Department of Medical Oncology, Papageorgiou Hospital, Aristotle University of Thessaloniki School of Medicine, Thessaloniki, Greece.
[40] Molecular Diagnostics Laboratory, INRASTES, National Centre for Scientific Research Demokritos, Aghia Paraskevi Attikis, Athens, GREECE
[41] Dana-Farber Cancer Institute, Boston, MA, USA
[42] Molecular Genetics of Breast Cancer, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany
[43] Clinical Genetics Department, St Michael's Hospital, Bristol, UK
[44] North West Thames Regional Genetics Service, Kennedy-Galton Centre, Harrow, UK
[45] Department of Clinical Genetics, Royal Devon & Exeter Hospital, Exeter, UK
[46] Cheshire & Merseyside Clinical Genetics Service, Liverpool Women's NHS Foundation Trust, Liverpool, UK
[47] Genetic Medicine, Manchester Academic Health Sciences Centre, Central Manchester University Hospitals NHS Foundation Trust, Manchester, UK
[48] Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, UK
[49] University of Southampton Faculty of Medicine, Southampton University Hospitals NHS Trust, Southampton, UK
[50] Institute of Genetic Medicine, Centre for Life, Newcastle Upon Tyne Hospitals NHS Trust, Newcastle upon Tyne, UK
[51] Sheffield Clinical Genetics Service, Sheffield Children's Hospital, Sheffield, UK
[52] Yorkshire Regional Genetics Service, Leeds, UK
[53] Leicestershire Clinical Genetics Service, University Hospitals of Leicester NHS Trust, UK
[54] Oxford Regional Genetics Service, Churchill Hospital, Oxford, UK
[55] Clinical Genetics, Guy's and St. Thomas' NHS Foundation Trust, London, UK
[56] North East Thames Regional Genetics Service, Great Ormond Street Hospital for Children NHS Trust, London, UK
[57] Academic Unit of Clinical and Molecular Oncology, Trinity College Dublin and St James's Hospital, Dublin, Eire
[58] Department of Clinical Genetics, East Anglian Regional Genetics Service, Addenbrookes Hospital, Cambridge, UK
[59] All Wales Medical Genetics Services, University Hospital of Wales, Cardiff, UK
[60] South East of Scotland Regional Genetics Service, Western General Hospital, Edinburgh, UK
[61] Centre for Cancer Research and Cell Biology, Queens University of Belfast, Department of Medical Genetics, Belfast HSC Trust, Belfast, UK
[62] Oncogenetics Team, The Institute of Cancer Research and Royal Marsden NHS Foundation Trust, UK
[63] Ferguson-Smith Centre for Clinical Genetics, Yorkhill Hospitals, Glasgow, UK
[64] Medical Genetics Unit, St George's, University of London, UK
[65] West Midlands Regional Genetics Service, Birmingham Women's Hospital Healthcare NHS Trust, Edgbaston, Birmingham, UK
[66] Department of Pathology and Laboratory Medicine, University of Kansas Medical Center, Kansas City, KS, USA
[67] Lombardi Comprehensive Cancer Center, Georgetown University, Washington DC, USA
[68] Center for Medical Genetics, Ghent University, Ghent, Belgium
[69] Department of Gynaecology and Obstetrics, Division of Tumor Genetics, Klinikum rechts der Isar, Technical University Munich, Germany
[70] Center of Familial Breast and Ovarian Cancer, Department of Medical Genetics, Institute of Human Genetics, University Würzburg, Germany

[71] Center for Hereditary Breast and Ovarian Cancer, Medical Faculty, University Hospital Cologne, Germany; Center for Integrated Oncology (CIO), Medical Faculty, University Hospital Cologne, Germany; Center for Molecular Medicine Cologne (CMMC), University of Cologne, Germany
[72] Institute of Human Genetics, Department of Human Genetics, University Hospital Heidelberg, Germany
[73] Institute for Medical Informatics, Statistics and Epidemiology
[74] Department of Gynaecology and Obstetrics, University Hospital Düsseldorf, Heinrich-Heine University Düsseldorf, Germany
[75] Institute of Cell and Molecular Pathology, Hannover Medical School, Hannover, Germany
[76] Institute of Human Genetics, University Medical Center Schleswig-Holstein, Campus Kiel, Germany
[77] Department of Gynaecology and Obstetrics, University Hospital Carl Gustav Carus, Technical University Dresden, Germany
[78] Department of Gynecology and Obstetrics, University Medical Center Schleswig-Holstein, Campus Kiel, Germany
[79] Institute of Human Genetics, Campus Virchov Klinikum, Charite Berlin, Germany
[80] Center for Hereditary Breast and Ovarian Cancer, Medical Faculty, University Hospital Cologne, Germany; Center for Integrated Oncology (CIO), Medical Faculty, University Hospital Cologne, Germany; Center for Molecular Medicine Cologne (CMMC), University of Cologne, Germany, on behalf of the German Consortium of Hereditary Breast and Ovarian Cncer (GC-HBOC)
[81] Institute of Human Genetics, University of Münster, Münster, Germany
[82] Department of Gynaecology and Obstetrics, University Hospital Ulm, Germany
[83] Institut Curie, Department of Tumour Biology, Paris, France
[84] Université Lyon 1, CNRS UMR5558, Lyon, France
[85] Unité de Prévention et d'Epidémiologie Génétique, Centre Léon Bérard, Lyon, France
[86] Department of Genetics, Centre Hospitalier Universitaire de Grenoble, Grenoble, France
[87] Institut Albert Bonniot, Université de Grenoble, Grenoble, France
[88] Laboratoire d'Oncogénétique, Hôpital René Huguenin/Institut Curie, Saint-Cloud, France
[89] GEMO study : National Cancer Genetics Network «UNICANCER Genetic Group», France
[90] Department of Genetics, Centre Hospitalier Universitaire de Grenoble, Grenoble, France
[91] Département d'Oncogénétique, Centre Jean Perrin, Université de Clermont-Ferrand, Clermont-Ferrand, France
[92] Université Lyon 1, CNRS UMR5558, Lyon, France
[93] Gynaecological Oncology, The University of Sydney Cancer Centre, Royal Prince Alfred Hospital, Sydney, AUSTRALIA
[94] Gynecologic Oncology Group, University of North Carolina at Chapel Hill, Chapel Hill, USA
[95] Gynecologic Oncology Group Statistical and Data Center, Roswell Park Cancer Institute, Buffalo, NY, USA
[96] Women & Infants Hospital, Providence, RI , US
[97] Molecular Oncology Laboratory, Hospital Clinico San Carlos, IdISSC, Madrid, Spain
[98] Department of Obstetrics and Gynecology, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland
[99] Department of Clinical Genetics, Helsinki University Central Hospital, Helsinki, Finland
[100] Department of Medical Oncology, Family Cancer Clinic, Erasmus University Medical Center, Rotterdam, The Netherlands
[101] Department of Clinical Genetics, Family Cancer Clinic, Erasmus University Medical Center, Rotterdam, The Netherlands
[102] Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands
[103] Department of Clinical Genetics, Academic Medical Center, Amsterdam, The Netherlands
[104] Department of Epidemiology, Netherlands Cancer Institute, Amsterdam, The Netherlands
[105] Family Cancer Clinic, Netherlands Cancer Institute, Amsterdam, The Netherlands
[106] Department of Clinical Genetics, VU University Medical Centre, Amsterdam, The Netherlands

[107] The Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON),Coordinating center: Netherlands Cancer Institute, Amsterdam, The Netherlands
[108] Department of Genetics, University Medical Center, Groningen University, Groningen, The Netherlands
[109] Department of Clinical Genetics, Maastricht University Medical Center, Maastricht, The Netherlands
[110] Department of Human Genetics & Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands
[111] Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands
[112] Department of Molecular Genetics, National Institute of Oncology, Budapest, Hungary
[113] Oncogenetics Group. University Hospital Vall d'Hebron, Vall d'Hebron Institute of Oncology (VHIO), Vall d'Hebron Research Institute (VHIR) and Universitat Autònoma de Barcelona; Barcelona, Spain
[114] Genetic Counseling Unit, Hereditary Cancer Program, IDIBELL-Catalan Institute of Oncology, Barcelona, Spain
[115] Molecular Diagnostic Unit, Hereditary Cancer Program, IDIBELL-Catalan Institute of Oncology, Barcelona, Spain
[116] Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland.
[117] Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland
[118] Landspitali University Hospital & University of Iceland School of Medicine, Reykjavik, Iceland
[119] Laboratoire de diagnostic génétique et Service d'Onco-hématologie, Hopitaux Universitaire de Strasbourg, CHRU Nouvel Hôpital Civil, STRASBOURG, France
[120] Department of Surgical Sciences, Oncology and Gastroenterology, Padua University and Immunology and Molecular Oncology Unit, Istituto Oncologico Veneto IOV - IRCCS, Padua, Italy
[121] Immunology and Molecular Oncology Unit, Istituto Oncologico Veneto IOV - IRCCS, Padua, Italy
[122] Department of Genetics, Portuguese Oncology Institute, Porto, Portugal, and Biomedical Sciences Institute (ICBAS), Porto University, Porto, Portugal
[123] Program in Cancer Genetics, Departments of Human Genetics and Oncology, McGill University, Montreal, Quebec, Canada
[124] Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[125] Health Sciences Research, Mayo Clinic, Scottsdale, AZ, USA
[126] National Human Genome Research Institute, National Institutes of Health, Building 31, Room 4B09,31 Center Drive, MSC 2152,9000 Rockville Pike, Bethesda, MD 20892-2152
[127] Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[128] Clinical Genetics Research Laboratory, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[129] Dept of OB/GYN and Comprehensive Cancer Center, Medical University of Vienna, Vienna, Austria
[130] Dept of OB/GYN and Comprehensive Cancer Center, Medical University of Austria, Austria
[131] Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD,USA
[132] Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, MD,USA
[133] Clalit National Israeli Cancer Control Center and Department of Community Medicine and Epidemiology, Carmel Medical Center and B. Rappaport Faculty of Medicine, Technion, 2 Horev St., Haifa, Israel
[134] N.N. Petrov Institute of Oncology, St.-Petersburg, Russia
[135] Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada; Department of Laboratory Medicine, and the Keenan Research Centre of the Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto, ON, Canada
[136] Ontario Cancer Genetics Network: Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario M5G 1X5, Cancer Care Ontario
[137] Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario M5G 1X5, Departments of Molecular Genetics and Laboratory Medicine and Pathobiology, University of Toronto, Ontario
[138] Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, Ontario, Canada

[139] Divison of Human Cancer Genetics, Departments of Internal Medicine and Molecular Virology, Immunology and Medical Genetics, Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA
[140] Section of Molecular Diagnostics, Department of Biochemistry, Aalborg University Hospital, Aalborg, Denmark
[141] Department of Clinical Genetics, Odense University Hospital, Odense C, Denmark
[142] Department of Clinical Genetics, Aarhus University Hospital, Aarhus N, Denmark
[143] Section of Genetic Oncology, Dept. of Laboratory Medicine, University and University Hospital of Pisa, Pisa, Italy
[144] Sheba Medical Center, Tel Aviv, Israel
[145] Institute of Oncology, Rivka Ziv Medical Center, Zefat, Israel
[146] Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden
[147] Department of Radiation Sciences, Oncology, Umeå University, Umea, Sweden
[148] Department of Oncology, Karolinska University Hospital, Stockholm, Sweden
[149] Department of Oncology, Lund University Hospital, Lund, Sweden
[150] Department of Immunology, Genetics and Pathology, Rudbeck Laboratory, Uppsala University, Uppsala, Sweden
[151] Center for Clinical Cancer Genetics and Global Health, University of Chicago Medical Center, Chicago, USA
[152] Department of Medicine and Genetics, University of California, San Francisco, USA
[153] Department of Preventive Medicine, Keck School of Medicine, University of Southern California, California, USA
[154] Department of Medicine, Abramson Cancer Center, Perelman School of Medicine at the University of Pennsylvania
[155] Department of Epidemiology and Biostatistics, Abramson Cancer Center, Perelman School of Medicine at the University of Pennsylvania
[156] University of Texas MD Anderson Cancer Center, Houston, TX, USA
[157] Familial Cancer Centre, Peter MacCallum Cancer Centre, Melbourne, AUSTRALIA, Sir Peter MacCallum, Department of Oncology, The University of Melbourne, Melbourne, Victoria, 3010, Australia
[158] Women's Cancer Program at the Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, USA
[159] Institut Curie, Department of Tumour Biology, Paris, France
[160] Institut Curie, INSERM U830, Paris, France
[161] Université Paris Descartes, Sorbonne Paris Cité, France
[162] Université Lyon 1, INCa-Synergie, Centre Léon Bérard, 28 rue Laennec, Lyon Cedex 08, France
[163] Department of Laboratory Medicine and Pathology, and Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[164] Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, Florida, USA
[165] Unité Mixte de Génétique Constitutionnelle des Cancers Fréquents, Hospices Civils de Lyon – Centre Léon Bérard, Lyon, France

FUNDING :

Higher level funding

The COGS project is funded through a European Commission's Seventh Framework Programme grant (agreement number 223175 - HEALTH-F2-2009-223175). The CIMBA data management and data analysis were supported by Cancer Research – UK grants C12292/A11174 and C1287/A10118. SH is supported by an NHMRC Program Grant to GCT.

Personal support

ACA is a Cancer Research-UK Senior Cancer Research Fellow (C12292/A11174). DFE is a Principal Research Fellow of Cancer Research UK. GC, MCS and IC are supported by the National Health and Medical Research Council. B.K. holds an American Cancer Society Early Detection Professorship (SIOP-06-258-01- COUN). Drs. Greene, Mai and Savage were supported by funding from the Intramural Research Program, NCI. OIO is an ACS Clinical Research Professor. JS is Chairholder of the Canada Research Chair in Oncogenetics.

Funding of constituent studies

BCFR was supported by grant UM1 CA164920 from the National Cancer Institute. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the BCFR; BFBOCC is partly supported by: Lithuania (BFBOCC-LT): Research Council of Lithuania grant LIG-07/2012; BFBOCC-LV is partly supported by LSC grant 10.0010.08 and in part by a grant from the ESF Nr.2009/0220/1DP/1.1.1.2.0/09/APIA/VIAA/016 and Liepaja's municipal council; BIDMC is supported by the Breast Cancer Research Foundation; BRCA-gene mutations and breast cancer in South African women (BMBSA) was supported by grants from the Cancer Association of South Africa (CANSA) to Elizabeth J. van Rensburg; BRICOH SLN was partially supported by the Morris and Horowitz Familes Endowed Professorship; CBCS was supported by the NEYE Foundation; CNIO was partially supported by Spanish Association against Cancer (AECC08), RTICC 06/0020/1060, FISPI08/1120, Mutua Madrileña Foundation (FMMA) and SAF2010-20493; City of Hope Clinical Cancer Genetics Community Network and the Hereditary Cancer Research Registry (COH-CCGCRN), supported in part by Award Number RC4CA153828 (PI: J. Weitzel) from the National Cancer Institute and the Office of the Director, National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health; CONSIT TEAM: Funds from Italian citizens who allocated the 5x1000 share of their tax payment in support of the Fondazione IRCCS Istituto Nazionale Tumori, according to Italian laws (INT-Institutional strategic projects '5x1000') to SM. Italian Association for Cancer Research (AIRC) to LO; DEMOKRITOS has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program of the General Secretariat for Research & Technology: ARISTEIA. Investing in knowledge society through the European Social Fund; DKFZ study was supported by the DKFZ; EMBRACE is supported by Cancer Research UK Grants C1287/A10118 and C1287/A11990. D. Gareth Evans and Fiona Lalloo are supported by an NIHR grant to the Biomedical Research Centre, Manchester. The Investigators at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust are supported by an NIHR grant to the Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust. Ros Eeles and Elizabeth Bancroft are supported by Cancer Research UK Grant C5047/A8385; KUMC: The authors acknowledge support from The University of Kansas Cancer Center (P30 CA168524) and the Kansas Bioscience Authority Eminent Scholar Program. A.K.G. was funded by 5U01CA113916, R01CA140323, and by the Chancellors Distinguished Chair in Biomedical Sciences Professorship; The German Consortium of Hereditary

Breast and Ovarian Cancer (GC-HBOC) is supported by the German Cancer Aid (grant no 109076, Rita K. Schmutzler) and by the Center for Molecular Medicine Cologne (CMMC)
GEMO was supported by the Ligue National Contre le Cancer; the Association “Le cancer du sein, parlons-en!” Award; and the Canadian Institutes of Health Research for the "CIHR Team in Familial Risks of Breast Cancer" program; G-FAST: Kim De Leeneer is supported by GOA grant BOF10/GOA/019 (Ghent University) and spearhead financing of Ghent University Hospital; GOG was supported by National Cancer Institute grants to the Gynecologic Oncology Group (GOG) Administrative Office and Tissue Bank (CA 27469), the GOG Statistical and Data Center (CA 37517), and GOG's Cancer Prevention and Control Committee (CA 101165); HCSC was supported by a grant RD12/00369/0006 and 12/00539 from ISCIII (Spain),partially supported by European Regional Development FEDER funds; HEBCS was financially supported by the Helsinki University Central Hospital Research Fund, Academy of Finland (266528), the Finnish Cancer Society and the Sigrid Juselius Foundation; HEBON is supported by the Dutch Cancer Society grants NKI1998-1854, NKI2004-3088, NKI2007-3756, the Netherlands Organization of Scientific Research grant NWO 91109024, the Pink Ribbon grant 110005 and the BBMRI grant NWO 184.021.007/CP46. HEBON thanks the registration teams of the Comprehensive Cancer Centre Netherlands and Comprehensive Centre South (together the Netherlands Cancer Registry) and PALGA (Dutch Pathology Registry) for part of the data collection; HRBCP is supported by The Hong Kong Hereditary Breast Cancer Family Registry and the Dr. Ellen Li Charitable Foundation, Hong Kong; HUNBOCS: Hungarian Breast and Ovarian Cancer Study was supported by Hungarian Research Grants KTIA-OTKA CK-80745 and OTKA K-112228; ICO: Contract grant sponsor: Asociación Española Contra el Cáncer; Spanish Health Research Foundation; Ramón Areces Foundation; Carlos III Health Institute; Catalan Health Institute; and Autonomous Government of Catalonia. Contract grant numbers: ISCIIIRETIC RD06/0020/1051, PI09/02483, PI10/01422, PI10/00748, PI13/00285, PI13/00189 2009SGR290 and 2009SGR283; IHCC was supported by Grant PBZ_KBN_122/P05/2004; ILUH was supported by the Icelandic Association “Walking for Breast Cancer Research” and by the Landspitali University Hospital Research Fund; INHERIT was supported by the Canadian Institutes of Health Research for the “CIHR Team in Familial Risks of Breast Cancer” program, the Canadian Breast Cancer Research Alliance-grant #019511 and the Ministry of Economic Development, Innovation and Export Trade – grant # PSR-SIIRI-701; IOVHBOCS is supported by Ministero della Salute and “5x1000” Istituto Oncologico Veneto grant; IPOBCS was in part supported by Liga Portuguesa Contra o Cancro; kConFab is supported by a grant from the National Breast Cancer Foundation, and previously by the National Health and Medical Research Council (NHMRC), the Queensland Cancer Fund, the Cancer Councils of New South Wales, Victoria, Tasmania and South Australia, and the Cancer Foundation of Western Australia; MAYO is supported by NIH grants CA116167, CA128978 and CA176785, an NCI Specialized Program of Research Excellence (SPORE) in Breast Cancer (CA116201), a U.S. Department of Defence Ovarian Cancer Idea award (W81XWH-10-1-0341), a grant from the Breast Cancer Research Foundation, a generous gift from the David F. and Margaret T. Grohne Family Foundation and the Ting Tsung and Wei Fong Chao Foundation; MCGILL : Jewish General Hospital Weekend to End Breast Cancer, Quebec Ministry of Economic Development, Innovation and Export Trade; MODSQUAD was supported by MH CZ - DRO (MMCI, 00209805) and by the European Regional Development Fund and the State Budget of the Czech Republic (RECAMO, CZ.1.05/2.1.00/03.0101) to LF, and by Charles University in Prague project UNCE204024 (MZ); MSKCC is supported by grants from the Breast Cancer Research Foundation and Robert and Kate Niehaus Clinical Cancer Genetics Initiative; NAROD: 1R01 CA149429-01; NCI: The research of Drs. MH Greene and PL Mai was supported by the Intramural Research Program of the US National Cancer Institute, NIH, and by support services contracts NO2-CP-11019-50 and N02-CP-65504 with Westat, Inc, Rockville, MD; NICCC is supported by Clalit Health Services in Israel. Some of its activities are supported by the Israel Cancer Association and the Breast Cancer Research Foundation (BCRF), NY.; NNPIO has been supported by the Russian Federation for Basic Research (grants 11-04-00227, 12-04-00928 and 12-04-01490) and the Federal Agency for Science and Innovations, Russia (contract 02.740.11.0780); OSUCCG is supported by the Ohio State University Comprehensive Cancer Center; PBCS was supported by the ITT (Istituto

Toscano Tumori) grants 2011-2013; SMC was partially funded through a grant by the Isreal cancer association and the funding for the Israeli Inherited breast cancer consortium; SWE-BRCA collaborators are supported by the Swedish Cancer Society; UCHICAGO is supported by NCI Specialized Program of Research Excellence (SPORE) in Breast Cancer (CA125183), R01 CA142996, 1U01CA161032 and by the Ralph and Marion Falk Medical Research Trust, the Entertainment Industry Fund National Women's Cancer Research Alliance and the Breast Cancer research Foundation; UCLA: Jonsson Comprehensive Cancer Center Foundation; Breast Cancer Research Foundation; UCSF Cancer Risk Program and Helen Diller Family Comprehensive Cancer Center; UKFOCR was supported by a project grant from CRUK to Paul Pharoah; UPENN: National Institutes of Health (NIH) (R01-CA102776 and R01-CA083855; Breast Cancer Research Foundation; Susan G. Komen Foundation for the cure, Basser Research Center for BRCA; VFCTG: Victorian Cancer Agency, Cancer Australia, National Breast Cancer Foundation; The Women's Cancer Program (WCP) at the Samuel Oschin Comprehensive Cancer Institute is funded by the American Cancer Society Early Detection Professorship (SIOP-06-258-01-COUN).

Correspondence To: David G. Cox
Cancer Research Center of Lyon
Bâtiment Cheney C, 2ème, Centre Léon Bérard
28 rue Laënnec 69373 Lyon Cedex 8 France
david.cox@lyon.unicancer.fr
+33 4 78 78 59 12 (voice) +33 4 78 78 28 04 (fax)

The authors disclose no potential conflicts of interest.

Word Count: ***4438***
Number of tables: ***4***
Number of figures: ***2***

ABSTRACT

Individuals carrying pathogenic mutations in *BRCA1/2* genes have a high lifetime risk of breast cancer. BRCA1 and BRCA2 are involved in DNA double strand break repair, DNA alterations that can be caused by exposure to reactive oxygen species, a main source of which are mitochondria. Mitochondrial genome variations affect electron transport chain efficiency and ROS production. Individuals from different mitochondrial haplogroups differ in their metabolism and sensitivity to oxidative stress. Variability in mitochondrial genetic background can alter ROS production, leading to cancer risk. Here we test the hypothesis that mitochondrial haplogroups modify breast cancer risk in *BRCA1/2* mutation carriers. We genotyped 22214 (11421 affected, 10793 unaffected) mutation carriers belonging to the Consortium of Investigators of Modifiers of *BRCA1/2* for 129 mitochondrial SNPs using the iCOGS array. Haplogroup inference and association detection were performed using a phylogenetic approach. ALTree was applied to explore the reference mitochondrial evolutionary tree and detect subclades enriched for affected or unaffected individuals. We discovered that subclade T1a1 was less enriched with affected *BRCA2* mutation carriers than the rest of clade T, (HR=0.55 (95% CI 0.34-0.88, p-value=0.01). Compared with the most frequent haplogroup in the general population *i.e.* H and T clade, the T1a1 haplogroup has an HR=0.62 (95% CI=0.40-0.95, p-value=0.03). We also identified three potential susceptibility loci, including G13708A/rs28359178, which has demonstrated an inverse association with familial breast cancer risk. This study illustrates how original approaches like the phylogeny-based method we used can empower classical molecular epidemiological studies aimed at identifying association or risk modification effects.

Introduction

Breast cancer is a multifactorial disease, with genetic, life-style and environmental susceptibility factors. Approximately 15-20% of the familial aggregation of breast cancer is accounted for by mutations in high-penetrance susceptibility genes (1),(2),(3) such as *BRCA1*, *BRCA2*. Pathogenic mutations in *BRCA1* and *BRCA2* confer lifetime breast cancer risk of 60%-85% (4),(5) and 40%-85%(4),(5), respectively. Other genomic variations (for example in genes encoding proteins interacting with BRCA1 and BRCA2) have been identified as modifiers of breast cancer risk and increase or decrease the risk initially conferred by *BRCA1* or *BRCA2* mutation(6).

BRCA1 and *BRCA2* are involved in DNA repair mechanisms, including double-strand break (DSB) repair by homologous recombination(7),(8) . DSB are considered to be one of the most deleterious forms of DNA damage because the integrity of both DNA strands is compromised simultaneously. These breaks can lead to genomic instability resulting in translocations, deletions, duplications, or mutations when not correctly repaired(9). Reactive Oxygen Species (ROS) are one of the main causes of DSBs, along with exposure to ionizing radiation, various chemical agents, and ultraviolet light (10).

ROS are naturally occurring chemical derivative of metabolism. Elevated levels of ROS and down-regulation of ROS scavengers and/or antioxidant enzymes can lead to oxidative stress, which is associated with a number of human diseases, including various cancers (11). The electron transport chain process, which takes place in the mitochondria, generates the majority of ROS in human cells. Variations in the mitochondrial genome have been shown to be associated with metabolic phenotypes and oxidative stress markers (12).

The human mitochondrial genome (mtDNA) has undergone a large number of mutations that have segregated during evolution. Those changes are now used to define mitochondrial haplogroups. Some of these changes slightly modify metabolic performance and energy production; thus, all haplogroups do not have identical metabolic capacities(13). It has been hypothesized that the geographic

distribution of mitochondrial haplogroups results from selection of metabolic capacities mainly driven by adaptation to climate and nutrition(14).

Mitochondrial haplogroups have been associated with diverse multifactorial diseases, such as Alzheimer's disease(15), hypertrophic cardiomyopathy(16) or age-related macular degeneration(17). Variations in mtDNA have also been linked to several types of cancer, such as gastric cancer(18) or renal cell carcinoma (19). Interestingly, variations in mtDNA have been linked to several types of female cancers: endometrial (20), ovarian(21), and breast cancer (22),(23).

The Collaborative Oncological Gene-environment Study(24) (COGS) is a European project designed to improve understanding of genetic susceptibility to breast, ovarian and prostate cancer. This project involves several consortia, the Breast Cancer Association Consortium (BCAC)(25) , the Ovarian Cancer Association Consortium (OCAC)(26) , the Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL)(27) , and the Consortium of Investigators of Modifiers of *BRCA1/2* (CIMBA)(28). CIMBA is a collaborative group of researchers working on genetic modifiers of cancer risk in *BRCA1* and *BRCA2* mutation carriers. As part of the COGS project, more than 200,000 single nucleotide polymorphisms (SNPs) were genotyped for *BRCA1* and *BRCA2* mutation carriers on the iCOGS chip, including 129 mitochondrial polymorphisms. The iCOGS chip is a custom Illumina™ Infinium genotyping array, designed to test, in a cost-effective manner, genetic variants related to breast, ovarian and prostate cancers.

In this study, we explored mitochondrial haplogroups as potential modifiers of breast cancer risk in women carrying pathogenic *BRCA1* or *BRCA2* mutations. Our study includes females diagnosed with breast cancer and unaffected carriers belonging to the CIMBA. We used an original analytic phylogenetic-based approach implemented in a homemade algorithm and in the program ALTree to infer haplogroups and to detect associations between haplogroups and breast cancer risk.

Materials and methods

BRCA1 and BRCA2 mutation carriers

All analyses were conducted separately on CIMBA *BRCA1* and *BRCA2* mutation carriers (abbreviated *pop1* and *pop2,* respectively). Eligible female carriers were aged 18 years or older and had a pathogenic mutation in *BRCA1* and/or *BRCA2*. Women mutated on both *BRCA1* and *BRCA2* were included in downstream analyses. Data were available on year of birth, age at study recruitment, age at cancer diagnosis, *BRCA1* and *BRCA2* mutation description, and self-reported ethnicity. Supplementary specifications regarding inclusion profiles are available in Couch *et al.* (29) and Gaudet *et al.* (30). Final analyses included 7,432 breast cancer cases and 7,104 unaffected *BRCA1* mutation carriers, and 3,989 invasive breast cancer and 3,689 unaffected *BRCA2* mutation carriers. Women with ovarian cancer history were not excluded from analyses, and represent 15% and 7% of *BRCA1* and *BRCA2* mutation carriers respectively.

Genotyping and Quality Filtering

Genotyping was conducted using the iCOGS custom Illumina Infinium array. Genotypes were called using Illumina's proprietary GenCall algorithm. Genotyping and quality filtering were described previously(29)-(30). Initially, 129 mitochondrial SNPs were genotyped for both *BRCA1* and *BRCA2* mutation carriers. SNPs fulfilling the following criteria were excluded from downstream analyses: monoallelic SNPs (minor allele frequency equals 0), SNPs with more than 5 % data missing, annotated as triallelic, or having probes cross-matching with the nuclear genome. Heterozygous genotypes were removed from analyses, and we further filtered out SNPs having more than 5 % of heterozygous calls, to limit potential for heteroplasmy affecting our results. We also did not retain SNPs representing private mutations. These mutations are rare, often restricted to a few families, and not sufficiently

prevalent in the general population to be included in the reference mitochondrial evolutionary tree (see below). This last step of filtration yielded 93 and 92 SNPs for the *pop1* and *pop2* analyses, respectively (see Supplementary Table S1 in Supplementary Material 1). Only individuals with fully defined haplotypes, *i.e.* non-missing genotypes for the 93 and 92 SNPs selected for *pop1* and *pop2* respectively were included in downstream analyses (14,536 and 7,678 individuals respectively).

Mitochondrial genome evolution and haplogroup definition

Analyses were based on the theoretical reconstructed phylogenetic tree of the mitochondrial genome (mtTree) known as PhyloTree(31) (v.15). The mtTree is rooted by the Reconstructed Sapiens Reference Sequence (RSRS). RSRS has been identified as the most likely candidate to root the mtTree by refining human mitochondrial phylogeny by parsimony (32). Each haplogroup in mtTree is defined by the set of mitochondrial genome SNPs that have segregated in RSRS until today in the mitochondrial genome. Each haplogroup is fully characterized by the 16569 bp sequence resulting from the application of all the substitutions that are encoded by the corresponding SNPs in RSRS sequence.

Haplogroups imputation

The phylogenetic approach used to infer haplogroups is described in Figure 1. Mitochondrial genome sequences can be reconstructed at each node of mtTree, given the substitutions that have segregated in RSRS. Each haplogroup therefore has a corresponding full-length mitochondrial sequence. However, the full-length mitochondrial sequence is not available in the data, since the iCOGs platform captured only 93 and 92 SNPs for pop1 and pop2 respectively. Thus, for each of the 7864 nodes of the phylogenetic tree, the corresponding short haplotype, i.e. the full-length sequence restricted to available loci was defined. Some of the short haplotypes are unique, and they can be matched with their corresponding haplogroup directly. However, most of the time, given the small number of SNPs

analyzed, several haplogroups correspond to the same short haplotype. Consequently, a unique haplogroup cannot confidently be assigned to each short haplotype. Therefore, each short haplotype was assigned the most recent common ancestor of all the haplogroups that share the same short haplotype. Once this matching was done, short haplotypes were reconstructed in the same way for each individual in our dataset, and assigned the corresponding haplogroup. Accuracy of the method used was assessed by being applied on a set of 630 mitochondrial genome sequences of known European and Caucasian haplogroups (See Supplementary Material 2)

Association Detection

This phylogenetic approach is based on the identification of subclades in the reference phylogenetic tree of the mitochondrial genome differentially enriched for cases and unaffected controls compared with neighboring subclades. We used ALTree(33)·(34) to perform association testing. ALTree - Association detection and Localization of susceptibility sites using haplotype phylogenetic Trees – is an algorithm performing nested homogeneity tests comparing distributions of affected and unaffected individuals in the different clades of a given phylogenetic tree. There are as many tests performed as levels in the phylogenetic tree. The p-value at each level of the tree is obtained by a permutation procedure in which 1000 permutations are performed. A procedure to correct for multiple testing adapted to nested tests (35) is implemented in ALTree. We aimed to detect a global association, so only the most significant p-value obtained for all tests performed on one tree is corrected.

Handling genetic dependency

ALTree performs homogeneity tests to detect differences in enrichment or depletion of affected or unaffected individuals between clades in the phylogenetic tree. This kind of test can only be performed on independent data. However because in the CIMBA dataset, some individuals belong to the same family, we constructed datasets with genetically independent data by randomly selecting

one individual among all those belonging to the same family and sharing the same short haplotype. To take into account the full variability of our data, we resampled one thousand times. Results of the analyses pipeline are obtained for each resampling independently, and then averaged over the one thousand re-samplings to obtain final results.

Character reconstruction at ancestral nodes

Before the ALTree localization algorithm was launched, ancestral sequences were reconstructed at each internal tree node*, i.e.* short haplotypes were inferred with maximum likelihood at all nodes that were not leaves. We used the software PAML(36) to perform the reconstruction at ancestral nodes using a maximum likelihood method. The phylogeny model used was the General Time-Reversible model (GTR or REV).

Localization of susceptibility sites

ALTree also includes an algorithm to identify which sites are the most likely ones to be involved in the association detected. For each short haplotype observed, the ALTree add-on *altree-add-S* will add to the short haplotype sequence a supplementary character called *S*, which represents the disease status associated to this short haplotype: are individuals carrying this short haplotype more often affected or unaffected? *S* is calculated based on the affected and unaffected counts, the relative proportion of affected and unaffected in the whole dataset, and sensibility parameter ε. ε was set to its default value, which is 1. After *S* character computation, haplotypes including character *S* are reconstructed at ancestral nodes. Susceptibility site localization is achieved with ALTree by computing a correlated evolution index calculated between each change of each site and the changes of the character *S*, in the two possible directions of change. The site(s) whose evolution is the most correlated to the character *S* is the most likely susceptibility site.

Selected subclades

Analyses were carried out on the full evolutionary tree. However, the more haplogroups there are at each level, the less statistical power homogeneity tests have. Therefore analyses were also applied to subclades extracted from the tree. Subclades were defined using counts of individuals in each haplogroup of the clade in order to maximize statistical power. Chosen subclades and corresponding affected and unaffected counts are presented in Table 1.

Statistical Analysis

We quantified the effect associated with enrichment discovered by applying ALTree by building a weighted Cox regression in which the outcome variable is the status (affected or non-affected) and the explicative variable is the inferred haplogroup. Analyses were stratified by country. Data were restricted to the clades of interest. The uncertainty in haplogroup inference was not taken into account in the model. The weighting method used takes into account breast cancer incidence rate as a function of age(37) and the gene containing the observed pathogenic mutation, *i.e. BRCA1* or *BRCA2*. Familial dependency was handled by using a robust sandwich estimate of variance (R package *survival*, *cluster()* function).

RESULTS

Haplogroup imputation

Supplementary Table S3 (in Supplementary Material 3) recapitulates absolute and relative frequencies for each haplogroup imputed in *BRCA1* and *BRCA2* mutation carriers. For *BRCA1* mutation carriers, we reconstructed 489 distinct short haplotypes of 93 loci from the genotypes data. Only 162 of those 489 short haplotypes matched theoretical haplotypes reconstructed in the reference mitochondrial evolutionary tree. These 162 haplotypes represented 13315 / 14536 individuals. Thus, 91.6 % of *BRCA1* mutation carriers were successfully assigned a haplogroup. For *BRCA2* mutation carriers, we reconstructed 350 distinct short haplotypes of 92 loci from our genotypes data. Only 139 of those 350 short haplotypes matched theoretical haplotypes reconstructed in the reference mitochondrial evolutionary tree. These 139 haplotypes represented 6996 / 7678 individuals. Thus, 91.1 % of *BRCA2* mutation carriers were successfully assigned a haplogroup. Since more *BRCA1* than *BRCA2* mutation carriers were genotyped (14,536 vs. 7,678 individuals), we logically observe more distinct haplotypes in *pop1* than in *pop2* (489 vs. 350 haplotypes).

Accuracy of the main haplogroup inference method used was estimate at 82%, and reached 100% for haplogroups I, J, K, T, U, W, X. Given the set of SNPs we dispose of, our method has difficulties to differentiate between H and V haplogroups (See 2 in Supplementary Material 2).

Association results

For both populations of *BRCA1* or *BRCA2* mutation carriers, and for the full tree as for all selected subclades, we extracted the mean corrected p-value for association testing over all resamplings

performed (See Table 2). The only corrected p-value that remained significant was that obtained for subclade T (abbreviated T*) in the population of individuals of *BRCA2* mutation carriers (p= 0.04).

The phylogenetic tree of subclade T (see Figure 2a) contains only three levels, thus only three tests were performed within this clade. It is legitimate to interpret the associated non-corrected p-values (see Table 3). Only the p-value associated with the test performed at the first level of the tree is significant. By looking more closely at the mean frequencies of affected and unaffected individuals in the tree at this level (see Figure 2b), we conclude that subclade T1a1 is less enriched for affected carriers than the neighboring subclades T and T2.

Localization results

We performed a localization analysis with ALTree. The correlated evolution index for all non-monomorphic sites observed in short haplotype sequences of subclade T are displayed in Supplementary Table S4 (in Supplementary Material 4). The higher the correlated evolution index, the more likely it is that corresponding sites will be involved in the observed association. Three short haplotype sites - numbered 44, 57, and 72 - are clearly distinguishing themselves, with correlation index values of 0.390, 0.324 and 0.318 respectively, whereas all other sites correlation index values ranged from -0.270 to -0.101. Table 4 shows details for these three loci.

Effect quantification

The ALTree method is able to detect an association, but cannot to quantify the associated effect. We estimated the risk of breast cancer for individuals with the T1a1 haplogroup compared with individuals having another T subclade haplogroup in the population of *BRCA2* mutation carriers with a more classical statistical method, a weighted Cox regression. We found a breast cancer HR=0.55

(95% CI=0.34-0.88, p-value = 0.014). We also tested Haplogroup T1a1 compared with other T* haplogroups and the H haplogroup (the main haplogroup in the general population), and found a breast cancer HR=0.62 (95% CI=0.40-0.95, p-value=0.03).

Discussion

We employed an original phylogenetic analytic method coupled with more classical molecular epidemiologic analyses in order to detect mitochondrial haplogroups differentially enriched for affected *BRCA1/2* mutation carriers. We successfully inferred haplogroups for more than 90% of individuals in our dataset. After haplogroup imputation, the ALTree method identified T1a1 in the T clade as differentially enriched in affected *BRCA2* mutation carriers, whereas no enrichment difference was found for *BRCA1* mutation carriers. The T subclade is present in 4% of African populations compared to 11% in Caucasian and east-European populations (38). In our data, the T subclade represented 9.34 % of *BRCA1* mutation carriers, and 9.30% of *BRCA2* carriers. The ALTree method also identified three potential breast cancer susceptibility loci in mitochondrial genome.

In this study, we investigated to what extent mitochondrial genome variability modified breast cancer risk in individuals carrying pathogenic mutations in *BRCA1/2*. A large proportion of breast cancer heritability still remains unexplained today (39). Different methods exist to study genomic susceptibility to a disease, such as linkage analyses (which identified the *BRCA1* and *BRCA2* susceptibility genes) or Genome-Wide Association Studies (GWAS). However, classical linkage analysis cannot be applied to the haploid mitochondrial genome. Furthermore, commercial GWAS chips available do not adequately capture the majority of mitochondrial SNPs. A non-genome-wide and mitochondrial focused approach was required to explore how mitochondrial genome variability influences breast cancer risk. Here we have shown that *BRCA2* mutation carriers representing the subclade T1a1 have between 30 and 50% less risk of breast cancer than those representing other clades which, if validated, is a clinically meaningful risk reduction and may influence choice of risk management strategies.

The association we observed among *BRCA2* but not *BRCA1* mutation carriers may reveal a functional alteration that would be specific to mechanisms involving *BRCA2*-related BC. Today it is established

that *BRCA1*- and *BRCA2*-associated breast cancers are not phenotypically identical. These two types of tumors do not harbor the same gene expression profiles or copy number alterations (40). Both BRCA1 and BRCA2 are involved in DNA damage repair via homologous recombination of double strands breaks, but also of single strand breaks (SSB), a much more frequent DNA damage that can be caused by ROS exposure(41). Recently, Davis *et al.* showed that an alternative homologous recombination (aHR) repair mechanism, different from the canonical one active on DSBs, was occurring on SSBs(42). As single-strand annealing (SSA), another repair mechanism that joins flanking repeated sequences *in cis* of a DNA break, this alternative homologous recombination mechanism is RAD51- and BRCA2-dependantly down-regulated. aHR is stimulated by the down regulation of RAD51 or BRCA2 expression or activity. Furthermore, aHR requires the presence of BRCA1. Thus, aHR may be active in contexts in which canonical homologous recombination is inactive, as for *BRCA2* mutation carriers. aHR is more prone to contribute to loss of heterozygosity (LOH), a source of mutations potentially driving tumorigenesis. Thus, aHR usage could explain why more LOH were observed in tumors characterized by homologous recombination deficiency, harboring BRCA1/2 alterations in ovarian tumors and in breast and prostate cancer cell lines(43). The association we discovered would be concordant with the hypothesis that individuals of T1a1 haplogroup might be less prone to genomic instability driven by use of aHR as single strand breaks repair mechanism than other *BRCA2* mutation carriers, and why we do not observe the same association in BRCA1 mutation carriers.

Three main reasons could explain our inability to assign haplogroups to 9% of study participants. First, given the high mutation rate in the mitochondrial genome, observed combinations of mitochondrial SNPs might have appeared relatively recently in the general population, and the corresponding haplotypes might not yet be incorporated in PhyloTree. Secondly, only one genotyping error could lead to chimeric haplotypes that do not exist although, given the quality of our genotyping data, this is unlikely. Finally, the mitochondrial reference evolutionary tree PhyloTree is based on phylogeny reconstruction by parsimony, and for some subclades it might be suboptimal, especially for

haplogroups relying on few mitochondrial sequences, as is the case for African haplogroups(44). In case of uncertainty, the choice we made to assign the most recent common ancestor to the studied haplotype enables us to improve statistical power without introducing a bias in the detected association. For the association detected between T, T1* and T2* subclades, the haplogroup inference method used do not bias the counts of affected and unaffected individuals in these subclades. More details are presented in Supplementary Table S5 in Supplementary Material 5. Furthermore, based on the haplogroup inference with our method of 630 European and Caucasian mitochondrial genome sequences whose haplogroup is known, we successfully assigned the correct main and subhaplogroup of 100% of sequences belonging to T, T2*, and T1a1* haplogroups.

We quantified the effect corresponding to the detected association by using a more classical approach. We built a weighted Cox regression including inferred haplogroup as explicative variable. However, the uncertainty in haplogroup inference was not taken into account in this model. Nevertheless, based on haplogroup assignment and regrouping performed in clade T, affected and unaffected counts of individuals in this clade were not biased.

With only 129 loci genotyped over the 16,569 nucleotides composing the mitochondrial genome, we certainly do not explore the full variability of mitochondrial haplotypes. A characterization of individual mitochondrial genomes would require more complete data acquisition methods to be used, *e.g.*, next-generation sequencing. However, next-generation sequencing presents its own limits and challenges, because some regions of the mitochondrial genome are not easily mappable due to a high homology with the nuclear genome among other factors, and important bioinformatics treatment is necessary to overcome sequencing technology biases. Finally, even for a relatively short genome of 'only' 16569 bp, mitochondrial genome sequencing of more than 20,000 individuals would represent a major increase in cost relative to genotyping 129 SNPs.

Mueller *et al*. (45) have investigated functional differences between haplogroups T and H, the most frequent haplogroup in the general population. They generated cybrids by combining HEK293 cells (human embryonic kidney cell line) devoid of mitochondrial DNA with isolated thrombocytes of various haplogroups. Interestingly, T haplogroup cybrids were found to have a higher capability to deal with oxidative stress and to survive when challenged with hydrogen peroxide than H haplogroup cybrids. However, Amo et al. (46) showed no significant bioenergetic differences in mitochondria with H or T-haplogroup mtDNAs in a constant nuclear background. Lin *et al*. (47) also used cybrids to explore mitochondrial haplogroups sensitivity to oxidative stress but whereas Mueller's study was focused on European haplogroups, they showed that Asian haplogroup B4b was more sensitive to oxidative stress induced by H_2O_2 exposure than the remaining frequent Asian haplogroups, whereas N9a cells demonstrated better capacities of survival in situation of oxidative stress. Unfortunately, variants of the N9b haplogroup do not seem to overlap with those we have been able to identify on haplogroup T1a1. Nevertheless, these studies provide some mechanistic evidence that T haplogroup carriers may be less sensitive to oxidative stress and subsequent DNA damage, supporting our observation of an inverse association between a subclade of this haplogroup and breast cancer risk among *BRCA2* mutation carriers.

ALTree identified T9899C, G11812A/rs41544217, and G13708A/rs28359178 as three potential susceptibility sites for the discovered association. G13708A is also known for being a secondary mutation for Leber's hereditary optic neuropathy (LHON). Although the role of secondary mutations in LHON is still controversial, G13708A could be associated with impairment of the respiratory chain in this pathology. G13708A has also been described as a somatic mutation in a breast cancer tumor, whereas it was not present in adjacent normal tissue and in blood leucocytes (48). A high proportion of mitochondrial somatic tumor specific variants are also known mitochondrial polymorphisms, which is consistent with the hypothesis that tumor cells are prone to acquire the same mutations that segregate into mitochondrial genome by selective adaptation when humans migrated out of Africa and were confronted to new environments(49). Interestingly, the germline variant G13708A has

already been shown to be inversely associated with familial breast cancer risk, with a breast cancer odds-ratio (OR)=0.47 (95% CI=0.24–0.92)(50).

This and our results suggest that mitochondrial haplogroups T1a1 modifies the individuals breast cancer risk. Further investigation of the biological mechanism behind the associations we observed may further reinforce the hypothesis that the mitochondrial genome is influential in breast cancer risk, particularly among carriers of *BRCA2* mutations and, if validated, is of a level to influence cancer risk management choices.

ACKOWLEDGMENTS

COGS

This study would not have been possible without the contributions of the following: Per Hall (COGS); Kyriaki Michailidou, Manjeet K. Bolla, Qin Wang (BCAC); Rosalind A. Eeles, Ali Amin Al Olama, Zsofia Kote-Jarai, Sara Benlloch (PRACTICAL); Alison M. Dunning, Craig Luccarini, Michael Lush and the staff of the Centre for Cancer Genetic Epidemiology; Simard and Daniel C. Tessier, Francois Bacot, Daniel Vincent, Sylvie LaBoissière and Frederic Robidoux and the staff of the McGill University and Génome Québec Innovation Centre; and Julie M. Cunningham, Sharon A. Windebank, Christopher A. Hilker, Jeffrey Meyer and the staff of Mayo Clinic Genotyping Core Facility;

CIMBA

Maggie Angelakos, Judi Maskiell, Gillian Dite, Helen Tsimiklis; members and participants in the New York site of the Breast Cancer Family Registry; members and participants in the Ontario Familial Breast Cancer Registry for their contributions to the study; Vilius Rudaitis, Laimonas Griškevičius, Drs Janis Eglitis, Anna Krilova and Aivars Stengrevics; the families who contribute to the BMBSA study;Chun Ding, Linda Steele; Alicia Barroso, Rosario Alonso, Guillermo Pita, Alessandra Viel and Lara della Puppa of the Centro di Riferimento Oncologico, IRCCS, Aviano (PN), Italy; Laura Papi of the University of Florence, Florence, Italy; Monica Barile of the Istituto Europeo di Oncologia, Milan, Italy; Liliana Varesco of the IRCCS AOU San Martino - IST Istituto Nazionale per la Ricerca sul Cancro, Genoa, Italy; Stefania Tommasi, Brunella Pilato and Rossana Lambo of the Istituto Nazionale Tumori "Giovanni Paolo II" - Bari, Italy; Aline Martayan of the Istituto Nazionale Tumori Regina Elena, Rome, Italy ; Maria Grazia Tibiletti of the Ospedale di Circolo-Università dell'Insubria, Varese, Italy; and the personnnel of the Cogentech Cancer Genetic Test Laboratory, Milan, Italy, EMBRACE Collaborating Centres are: Coordinating Centre, Cambridge: Debra Frost, Steve Ellis, Elena Fineberg, Radka Platte. North of Scotland Regional Genetics Service, Aberdeen: Zosia Miedzybrodzka, Helen Gregory. Northern Ireland Regional Genetics Service, Belfast: Patrick Morrison, Lisa Jeffers. West Midlands Regional Clinical Genetics Service, Birmingham: Trevor Cole, Kai-ren Ong, Jonathan Hoffman. South West Regional Genetics Service, Bristol: Alan Donaldson, Margaret James. East Anglian Regional Genetics Service, Cambridge: Marc Tischkowitz, Joan Paterson, Amy Taylor. Medical Genetics Services for Wales, Cardiff: Alexandra Murray, Mark T. Rogers, Emma McCann. St James's Hospital, Dublin & National Centre for Medical Genetics, Dublin: M. John Kennedy, David Barton. South East of Scotland Regional Genetics Service, Edinburgh: Mary Porteous, Sarah Drummond. Peninsula Clinical Genetics Service, Exeter: Carole Brewer, Emma Kivuva, Anne Searle, Selina Goodman, Kathryn Hill. West of Scotland Regional Genetics Service, Glasgow: Rosemarie Davidson, Victoria Murday, Nicola Bradshaw, Lesley Snadden, Mark Longmuir, Catherine Watt, Sarah Gibson, Eshika Haque, Ed Tobias, Alexis Duncan. South East Thames Regional Genetics Service, Guy's Hospital London: Louise Izatt, Chris Jacobs, Caroline Langman. North West Thames Regional Genetics Service, Harrow: Huw Dorkins. Leicestershire Clinical Genetics Service, Leicester: Julian Barwell. Yorkshire Regional Genetics Service, Leeds: Julian Adlard, Gemma Serra-Feliu. Cheshire & Merseyside Clinical Genetics Service, Liverpool: Ian Ellis, Catherine Houghton. Manchester Regional Genetics Service, Manchester: D Gareth Evans, Fiona Lalloo, Jane Taylor. North East Thames Regional Genetics Service, NE Thames, London: Lucy Side, Alison Male, Cheryl Berlin. Nottingham Centre for Medical Genetics, Nottingham: Jacqueline Eason, Rebecca Collier. Northern Clinical Genetics Service, Newcastle: Fiona Douglas, Oonagh Claber, Irene Jobson. Oxford Regional Genetics Service, Oxford: Lisa Walker, Diane McLeod, Dorothy Halliday, Sarah Durell, Barbara Stayner. The Institute of Cancer Research and Royal Marsden NHS Foundation Trust: Ros Eeles, Susan Shanley, Nazneen Rahman, Richard Houlston, Elizabeth Bancroft, Elizabeth Page, Audrey Ardern-Jones, Kelly Kohut, Jennifer Wiggins, Elena Castro, Emma Killick, Sue Martin, Gillian Rea, Anjana Kulkarni. North Trent Clinical Genetics Service, Sheffield: Jackie Cook, Oliver Quarrell, Cathryn Bardsley. South West Thames Regional Genetics Service, London: Shirley Hodgson, Sheila Goff, Glen Brice, Lizzie Winchester, Charlotte

Eddy, Vishakha Tripathi, Virginia Attard, Anna Lehmann. Wessex Clinical Genetics Service, Princess Anne Hospital, Southampton: Diana Eccles, Anneke Lucassen, Gillian Crawford, Donna McBride, Sarah Smalley; Ms. JoEllen Weaver and Dr. Betsy Bove for their technical support; Genetic Modifiers of Cancer Risk in BRCA1/2 Mutation Carriers (GEMO) study : National Cancer Genetics Network UNICANCER Genetic Group, France. GEMO Collaborating Centers are: Coordinating Centres, Unité Mixte de Génétique Constitutionnelle des Cancers Fréquents, Hospices Civils de Lyon - Centre Léon Bérard, & Equipe «Génétique du cancer du sein», Centre de Recherche en Cancérologie de Lyon: Olga Sinilnikova, Sylvie Mazoyer, Francesca Damiola, Laure Barjhoux, Carole Verny-Pierre, Alain Calender, Sophie Giraud, Mélanie Léone; and Service de Génétique Oncologique, Institut Curie, Paris: Dominique Stoppa-Lyonnet, Marion Gauthier-Villars, Bruno Buecher, Claude Houdayer, Virginie Moncoutier, Muriel Belotti, Carole Tirapo, Antoine de Pauw. Institut Gustave Roussy, Villejuif: Brigitte Bressac-de-Paillerets, Olivier Caron. Centre Jean Perrin, Clermont–Ferrand: Yves-Jean Bignon, Nancy Uhrhammer. Centre Léon Bérard, Lyon: Christine Lasset, Valérie Bonadona, Sandrine Handallou. Centre François Baclesse, Caen: Agnès Hardouin, Pascaline Berthet. Institut Paoli Calmettes, Marseille: Hagay Sobol, Violaine Bourdon, Tetsuro Noguchi, Audrey Remenieras, François Eisinger. CHU Arnaud-de-Villeneuve, Montpellier: Isabelle Coupier, Pascal Pujol. Centre Oscar Lambret, Lille: Jean-Philippe Peyrat, Joëlle Fournier, Françoise Révillion, Philippe Vennin, Claude Adenis. Hôpital René Huguenin/Institut Curie, St Cloud: Etienne Rouleau, Rosette Lidereau, Liliane Demange, Catherine Nogues. Centre Paul Strauss, Strasbourg: Danièle Muller, Jean-Pierre Fricker. Institut Bergonié, Bordeaux: Emmanuelle Barouk-Simonet, Françoise Bonnet, Virginie Bubien, Nicolas Sevenet, Michel Longy. Institut Claudius Regaud, Toulouse: Christine Toulas, Rosine Guimbaud, Laurence Gladieff, Viviane Feillel. CHU Grenoble: Dominique Leroux, Hélène Dreyfus, Christine Rebischung, Magalie Peysselon. CHU Dijon: Fanny Coron, Laurence Faivre. CHU St-Etienne: Fabienne Prieur, Marine Lebrun, Caroline Kientz. Hôtel Dieu Centre Hospitalier, Chambéry: Sandra Fert Ferrer. Centre Antoine Lacassagne, Nice: Marc Frénay. CHU Limoges: Laurence Vénat-Bouvet. CHU Nantes: Capucine Delnatte. CHU Bretonneau, Tours: Isabelle Mortemousque. Groupe Hospitalier Pitié-Salpétrière, Paris: Florence Coulet, Chrystelle Colas, Florent Soubrier. CHU Vandoeuvre-les-Nancy : Johanna Sokolowska, Myriam Bronner. CHU Besançon: Marie-Agnès Collonge-Rame, Alexandre Damette. Creighton University, Omaha, USA: Henry T.Lynch, Carrie L.Snyder; the technical support of Ilse Coene en Brecht Crombez; the investigators of the Australia New Zealand Gynaecological Oncology Group (ANZGOG). We acknowledge Alicia Tosar for her technical assistance; Taru A. Muranen, Drs. Carl Blomqvist and Kirsimari Aaltonen and RNs Irja Erkkilä and Virpi Palola for their help with the HEBCS data and samples; The Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON) consists of the following Collaborating Centers: Coordinating center: Netherlands Cancer Institute, Amsterdam, NL: M.A. Rookus, F.B.L. Hogervorst, F.E. van Leeuwen, S. Verhoef, M.K. Schmidt, J.L. de Lange, R. Wijnands; Erasmus Medical Center, Rotterdam, NL: J.M. Collée, A.M.W. van den Ouweland, M.J. Hooning, C. Seynaeve, C.H.M. van Deurzen, I.M. Obdeijn; Leiden University Medical Center, NL: C.J. van Asperen, J.T. Wijnen, R.A.E.M. Tollenaar, P. Devilee, T.C.T.E.F. van Cronenburg; Radboud University Nijmegen Medical Center, NL: C.M. Kets, A.R. Mensenkamp; University Medical Center Utrecht, NL: M.G.E.M. Ausems, R.B. van der Luijt; Amsterdam Medical Center, NL: C.M. Aalfs, T.A.M. van Os; VU University Medical Center, Amsterdam, NL: J.J.P. Gille, Q. Waisfisz, H.E.J. Meijers-Heijboer; University Hospital Maastricht, NL: E.B. Gómez-Garcia, M.J. Blok; University Medical Center Groningen, NL: J.C. Oosterwijk, A.H. van der Hout, M.J. Mourits, G.H. de Bock; The Netherlands Foundation for the detection of hereditary tumours, Leiden, NL: H.F. Vasen; The Netherlands Cancer Registry: S. Siesling; The Dutch Pathology Registry (PALGA): L.I.H. Overbeek; Hong Kong Sanatoriuma and Hospital for their continual support; Janos Papp, Tibor Vaszko, Aniko Bozsik, Timea Pocza, Judit Franko, Maria Balogh, Gabriella Domokos, Judit Ferenczi (Department of Molecular Genetics, National Institute of Oncology, Budapest, Hungary) and the clinicians and patients for their contributions to this study; the Oncogenetics Group, and the High Risk and Cancer Prevention Unit of the University Hospital Vall d'Hebron led by Dr. J. Balmaña; the ICO Hereditary Cancer Program team team led by Dr. Gabriel Capella; Dr Martine Dumont, Martine Tranchant for sample management and skillful technical assistance. J.S. and P.S. were part of the QC and Genotyping coordinating group of iCOGS (BCAC and

CIMBA); Drs. Ana Peixoto, Catarina Santos, Patrícia Rocha and Pedro Pinto for their skillful contribution to the study; Heather Thorne, Eveline Niedermayr, all the kConFab research nurses and staff, the heads and staff of the Family Cancer Clinics, and the Clinical Follow Up Study (which has received funding from the NHMRC, the National Breast Cancer Foundation, Cancer Australia, and the National Institute of Health (USA)) for their contributions to this resource, and the many families who contribute to kConFab; Lenka Foretova and Eva Machackova (Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute and MF MU, Brno, Czech Republic); Michal Zikan, Petr Pohlreich and Zdenek Kleibl (Oncogynecologic Center and Department of Biochemistry and Experimental Oncology, First Faculty of Medicine, Charles University, Prague, Czech Republic);Anne Lincoln, Lauren Jacobs; the NICCC National Familial Cancer Consultation Service team led by Sara Dishon, the lab team led by Dr. Flavio Lejbkowicz, and the research field operations team led by Dr. Mila Pinchev; members and participants in the Ontario Cancer Genetics Network for their contributions to the study; Leigha Senter, Kevin Sweet, Caroline Craven, and Michelle O'Conor were instrumental in accrual of study participants, ascertainment of medical records and database management; the OSU Human Genetics Sample Bank; the Meirav Comprehensive breast cancer center team at the Sheba Medical Center; Åke Borg, Håkan Olsson, Helena Jernström, Karin Henriksson, Katja Harbst, Maria Soller, Ulf Kristoffersson; from Gothenburg Sahlgrenska University Hospital: Anna Öfverholm, Margareta Nordling, Per Karlsson, Zakaria Einbeigi; from Stockholm and Karolinska University Hospital: Anna von Wachenfeldt, Annelie Liljegren, Annika Lindblom, Brita Arver, Gisela Barbany Bustinza, Johanna Rantala; from Umeå University Hospital: Beatrice Melin, Christina Edwinsdotter Ardnor, Monica Emanuelsson; from Uppsala University: Hans Ehrencrona, Maritta Hellström Pigg, Richard Rosenquist; from Linköping University Hospital: Marie Stenmark-Askmalm, Sigrun Liedgren; Cecilia Zvocec, Qun Niu, physicians, genetic counselors, research nurses and staff of the Cancer Risk Clinic for their contributions to this resource; Joyce Seldon MSGC and Lorna Kwan, MPH; Dr. Robert Nussbaum and the following genetic counselors: Beth Crawford, Kate Loranger, Julie Mak, Nicola Stewart, Robin Lee, Amie Blanco and Peggy Conrad; Ms. Salina Chan; Paul Pharoah, Simon Gayther, Susan Ramus, Carole Pye, Patricia Harrington and Eva Wozniak for their contributions towards the UKFOCR; Geoffrey Lindeman, Marion Harris, Martin Delatycki of the Victorian Familial Cancer Trials Group; Sarah Sawyer, Rebecca Driessen and Ella Thompson.

AUTHORS CONTRIBUTION

Conception and design: DGC GT.

Development of methodology: SB CBa VD.

Acquisition of data: LM, SHe, DB, ALe, JD, KBK, PS, MBT, WKC, DEG, SSB, RJ, LT, NT, CMD, EJvR, SLN, YCD, AMG, BE, FCN, TvOH, AO, JBe, RA, ES, JNW, MThe, PP, PR, VP, RDo, BB, BP, DZ, GSc, SMan, LV, GLC, LP, LO, DY, IK, JGa, UH, AD, ABr, CBr, CF, DGE, DF, DE, FDo, JCo, JA, JBa, LW, LI, LES, MJK, MTi, MTR, MEP, PJM, RP, RE, RDa, SHo, TCo, AKG, CI, KC, KDL, AM, AG, BW, CS, CE, DN, DS, HP, KK, KR, ND, NA, RV, RKS, SP, SW, AdP, CLe, CLas, DL, ER, FDa, GSC, HD, LB, LG, NU, VB, VS, YB, JCa, LVL, MP, PAD, MdlH, TCa, HN, KA, AJag, AMvdO, CMK, CMA, FEvL, FBH, HEM, JCO, KvR, MAR, PD, RBvdL, EO, OD, AT, CLaz, IB, JDV, AJak, GSu, JGr, JLu, KD, KJ, BAA, CM, AA, MM, MRT, ABS, WF, CO, NLi, VSP, CIS, ALinc, LJ, MC, MR, JV, ABe, AF, CFS, CR, DGK, GP, MTe, MHG, PLM, GR, EI, AMM, GG, ILA, ST, AET, ISP, MTho, TAK, UBJ, MAC, EF, JZ, YL, ALind, BM, BA, NLo, RR, OIO, RLN, SR, KLN, SMD, TRR, BKA, GM, BYK, JLe, SO, DSL, GT, JS, FJC, KO, DFE, GC, ACA, SMaz, CMP, OMS.

Analysis and interpretation of data: SB DGC ACA.

Writing the manuscript: SB DGC ACA SH ABS GC SLN AET ILA JCO VJ KO MTho GM.

References

1. Newman B, Austin MA, Lee M, King MC. Inheritance of human breast cancer: evidence for autosomal dominant transmission in high-risk families. Proc Natl Acad Sci U S A. 1988;85:3044–8.

2. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, et al. Linkage of early-onset familial breast cancer to chromosome 17q21. Science. 1990;250:1684–9.

3. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science. 1994;266:66–71.

4. Evans DG, Shenton A, Woodward E, Lalloo F, Howell A, Maher ER. Penetrance estimates for BRCA1 and BRCA2 based on genetic testing in a Clinical Cancer Genetics service setting: risks of breast/ovarian cancer quoted should reflect the cancer burden in the family. BMC Cancer. 2008;8:155.

5. Antoniou A, Pharoah PDP, Narod S, Risch HA, Eyfjord JE, Hopper JL, et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. Am J Hum Genet. 2003;72:1117–30.

6. Lynch HT, Snyder C, Casey MJ. Hereditary ovarian and breast cancer: what have we learned? Ann Oncol. 2013;24:viii83–viii95.

7. Caestecker KW, Van de Walle GR. The role of BRCA1 in DNA double-strand repair: past and present. Exp Cell Res. 2013;319:575–87.

8. Negritto C. Repairing Double-Strand DNA Breaks. Nat Educ. 2010;3:26.

9. Osorio A, Milne RL, Kuchenbaecker K, Vaclová T, Pita G, Alonso R, et al. DNA Glycosylases Involved in Base Excision Repair May Be Associated with Cancer Risk in BRCA1 and BRCA2 Mutation Carriers. PLoS Genet. 2014;10:e1004256.

10. Klaunig JE, Kamendulis LM, Hocevar BA. Oxidative Stress and Oxidative Damage in Carcinogenesis. Toxicol Pathol. 2010;38:96–109.

11. Loft S, Poulsen HE. Cancer risk and oxidative DNA damage in man. J Mol Med Berl. 1996;74:297–312.

12. Weng S-W, Lin T-K, Wang P-W, Chen S-D, Chuang Y-C, Liou C-W. Single nucleotide polymorphisms in the mitochondrial control region are associated with metabolic phenotypes and oxidative stress. Gene. 2013;531:370–6.

13. Kenney MC, Chwa M, Atilano SR, Falatoonzadeh P, Ramirez C, Malik D, et al. Molecular and bioenergetic differences between cells with African versus European inherited mitochondrial DNA haplogroups: Implications for population susceptibility to diseases. Biochim Biophys Acta BBA - Mol Basis Dis. 2014;1842:208–19.

14. Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, et al. Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci U S A. 2003;100:171–6.

15. Van der Walt JM, Dementieva YA, Martin ER, Scott WK, Nicodemus KK, Kroner CC, et al. Analysis of European mitochondrial haplogroups with Alzheimer disease risk. Neurosci Lett.

2004;365:28–32.

16. Castro MG, Huerta C, Reguero JR, Soto MI, Doménech E, Alvarez V, et al. Mitochondrial DNA haplogroups in Spanish patients with hypertrophic cardiomyopathy. Int J Cardiol. 2006;112:202–6.

17. Mueller EE, Schaier E, Brunner SM, Eder W, Mayr JA, Egger SF, et al. Mitochondrial Haplogroups and Control Region Polymorphisms in Age-Related Macular Degeneration: A Case-Control Study. PLoS ONE. 2012;7:e30874.

18. Wei L, Zhao Y, Guo T, Li P, Wu H, Xie H, et al. Association of mtDNA D-loop polymorphisms with risk of gastric cancer in Chinese population. Pathol Oncol Res POR. 2011;17:735–42.

19. Zhang J, Guo Z, Bai Y, Cui L, Zhang S, Xu J. Identification of sequence polymorphisms in the displacement loop region of mitochondrial DNA as a risk factor for renal cell carcinoma. Biomed Rep. 2013;1:563–6.

20. Liu VWS, Wang Y, Yang H-J, Tsang PCK, Ng T-Y, Wong L-C, et al. Mitochondrial DNA variant 16189T>C is associated with susceptibility to endometrial cancer. Hum Mutat. 2003;22:173–4.

21. Permuth-Wey J, Chen YA, Tsai Y-Y, Chen Z, Qu X, Lancaster JM, et al. Inherited Variants in Mitochondrial Biogenesis Genes May Influence Epithelial Ovarian Cancer Risk. Cancer Epidemiol Biomarkers Prev. 2011;20:1131–45.

22. Czarnecka AM, Krawczyk T, Zdrozny M, Lubiński J, Arnold RS, Kukwa W, et al. Mitochondrial NADH-dehydrogenase subunit 3 (ND3) polymorphism (A10398G) and sporadic breast cancer in Poland. Breast Cancer Res Treat. 2010;121:511–8.

23. Mims MP, Hayes TG, Zheng S, Leal SM, Frolov A, Ittmann MM, et al. Mitochondrial DNA G10398A polymorphism and invasive breast cancer in African-American women. Cancer Res. 2006;66:1880; author reply 1880–1881.

24. Bahcall O. COGS project and design of the iCOGS array. Nat Genet [Internet]. 2013 [cited 2014 Jan 13]; Available from: http://www.nature.com/icogs/primer/cogs-project-and-design-of-the-icogs-array/

25. Breast Cancer Association Consortium. Commonly studied single-nucleotide polymorphisms and breast cancer: results from the Breast Cancer Association Consortium. J Natl Cancer Inst. 2006;98:1382–96.

26. Berchuck A, Schildkraut JM, Pearce CL, Chenevix-Trench G, Pharoah PD. Role of Genetic Polymorphisms in Ovarian Cancer Susceptibility: Development of an International Ovarian Cancer Association Consortium. In: Coukos G, Berchuck A, Ozols R, editors. Ovarian Cancer [Internet]. Springer New York; 2008 [cited 2014 Feb 7]. page 53–67. Available from: http://link.springer.com.gate2.inist.fr/chapter/10.1007/978-0-387-68969-2_5

27. Kote-Jarai Z, Easton DF, Stanford JL, Ostrander EA, Schleutker J, Ingles SA, et al. Multiple novel prostate cancer predisposition loci confirmed by an international study: the PRACTICAL Consortium. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2008;17:2052–61.

28. Chenevix-Trench G, Milne RL, Antoniou AC, Couch FJ, Easton DF, Goldgar DE, et al. An international initiative to identify genetic modifiers of cancer risk in BRCA1 and BRCA2 mutation carriers: the Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA). Breast Cancer Res BCR. 2007;9:104.

29. Couch FJ, Wang X, McGuffog L, Lee A, Olswold C, Kuchenbaecker KB, et al. Genome-Wide Association Study in BRCA1 Mutation Carriers Identifies Novel Loci Associated with Breast and Ovarian Cancer Risk. PLoS Genet. 2013;9:e1003212.

30. Gaudet MM, Kuchenbaecker KB, Vijai J, Klein RJ, Kirchhoff T, McGuffog L, et al. Identification of a BRCA2-Specific Modifier Locus at 6p24 Related to Breast Cancer Risk. PLoS Genet. 2013;9:e1003173.

31. Van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat. 2009;30:E386–E394.

32. Behar DM, van Oven M, Rosset S, Metspalu M, Loogväli E-L, Silva NM, et al. A "Copernican" Reassessment of the Human Mitochondrial DNA Tree from its Root. Am J Hum Genet. 2012;90:675–84.

33. Bardel C, Danjean V, Génin E. ALTree: association detection and localization of susceptibility sites using haplotype phylogenetic trees. Bioinformatics. 2006;22:1402–3.

34. Bardel C, Danjean V, Morange P, Génin E, Darlu P. On the use of phylogeny-based tests to detect association between quantitative traits and haplotypes. Genet Epidemiol. 2009;33:729–39.

35. Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. Test. 2003;12:1–77.

36. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91.

37. Antoniou AC, Goldgar DE, Andrieu N, Chang-Claude J, Brohet R, Rookus MA, et al. A weighted cohort approach for analysing factors modifying disease risks in carriers of high-risk susceptibility genes. Genet Epidemiol. 2005;29:1–11.

38. MITOMAP : Haplogroups frequencies [Internet]. Available from: http://www.mitomap.org/bin/view.pl/MITOMAP/HaplogroupMarkers

39. Shuen AY, Foulkes WD. Inherited Mutations in Breast Cancer Genes—Risk and Response. J Mammary Gland Biol Neoplasia. 2011;16:3–15.

40. Jönsson G, Naylor TL, Vallon-Christersson J, Staaf J, Huang J, Ward MR, et al. Distinct genomic profiles in hereditary breast tumors identified by array-based comparative genomic hybridization. Cancer Res. 2005;65:7612–21.

41. Caldecott KW. Single-strand break repair and genetic disease. Nat Rev Genet. 2008;9:619–31.

42. Davis L, Maizels N. Homology-directed repair of DNA nicks via pathways distinct from canonical double-strand break repair. Proc Natl Acad Sci U S A. 2014;111:E924–932.

43. Abkevich V, Timms KM, Hennessy BT, Potter J, Carey MS, Meyer LA, et al. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. Br J Cancer. 2012;107:1776–82.

44. Bandelt H-J, Kloss-Brandstätter A, Richards MB, Yao Y-G, Logan I. The case for the continuing use of the revised Cambridge Reference Sequence (rCRS) and the standardization of notation in human mitochondrial DNA studies. J Hum Genet [Internet]. 2013 [cited 2014 Jan 10]; Available from: http://www.nature.com.gate2.inist.fr/jhg/journal/vaop/ncurrent/full/jhg2013120a.html

45. Mueller EE, Brunner SM, Mayr JA, Stanger O, Sperl W, Kofler B. Functional Differences between Mitochondrial Haplogroup T and Haplogroup H in HEK293 Cybrid Cells. PLoS ONE.

2012;7:e52367.

46. Amo T, Yadava N, Oh R, Nicholls DG, Brand MD. Experimental assessment of bioenergetic differences caused by the common European mitochondrial DNA haplogroups H and T. Gene. 2008;411:69–76.

47. Lin T-K, Lin H-Y, Chen S-D, Chuang Y-C, Chuang J-H, Wang P-W, et al. The Creation of Cybrids Harboring Mitochondrial Haplogroups in the Taiwanese Population of Ethnic Chinese Background: An Extensive *In Vitro* Tool for the Study of Mitochondrial Genomic Variations. Oxid Med Cell Longev [Internet]. 2012 [cited 2014 Jan 21];2012. Available from: http://www.hindawi.com/journals/omcl/2012/824275/abs/

48. Parrella P, Xiao Y, Fliss M, Sanchez-Cespedes M, Mazzarelli P, Rinaldi M, et al. Detection of Mitochondrial DNA Mutations in Primary Breast Cancer and Fine-Needle Aspirates. Cancer Res. 2001;61:7623–6.

49. Brandon M, Baldi P, Wallace DC. Mitochondrial mutations in cancer. Oncogene. 2006;25:4647–62.

50. Bai R-K, Leal SM, Covarrubias D, Liu A, Wong L-JC. Mitochondrial genetic background modifies breast cancer risk. Cancer Res. 2007;67:4687–94.

TABLES

Subclade	BRCA1 mutation carriers	BRCA2 mutation carriers
U8	1458	863
T	1243	651
J	1270	630
J1	1043	513
H	3706	1967
H1	582	337
U5	868	458
X1'2'3	221	103
K1a	608	364

Table 1: Counts of participants in selected subclades

Subclade	*pop1* corrected p-value	*pop2* corrected p-value
Full	0.8298671	0.680985
U8	0.1457532	0.6260519
T	0.2854815	**0.0402038**
J	0.7175275	0.1115694
J1	0.6214585	0.1491129
H	0.7474476	0.9302557
H1	0.2677572	0.8035485
U5	0.8293806	0.7474615
X1'2'3	0.4155115	0.6288012
K1a	0.1701149	0.1617493

Table 2 : Mean corrected p-value for association testing with ALTree

Level	Degrees of freedom	Mean of non-corrected p-value
1	2	0.02141039
2	6	0.14355900
3	8	0.22249700

Table 3 : Non-corrected p-values by level of phylogenetic tree for subclade T in *BRCA2* mutation carriers

Site	SNP Name	Position	Direction of change	Correlated evolution index	Major Allele	Minor Allele	MAF in *pop2*
44	MitoT9900C	9899	T → C	0.390	T	C	0.0163
57	rs41544217	11812	G →A	0.324	A	G	0.0709
72	rs28359178	13708	G →A	0.318	G	A	0.1106

Table 4: Description of loci identified as potential susceptibility sites by ALTree

LEGEND OF FIGURES

Figure 1: Simplified representation of the phylogenic method used to infer haplogroups. **a.** Full-length haplotypic sequences are reconstructed at each node of the reference tree. **b.** Haplotypes are then restricted to available loci. Sequences of the same color are identical. **c.** Unique short haplotypes are matched directly with the corresponding haplogroup. **d.** Sequences matching with several haplogroups are associated to their Most Recent common Ancestor haplogroup.

Figure 2: Phylogenetic tree of subclade T tested for association with ALTree **a.** Phylogenetic tree of subclade T with all observed haplogroups. A homogeneity test is performed at each level of the tree. **b.** First level of phylogenetic tree of subclade T. Averaged counts from resampling of affected and unaffected are indicated below each subclade, respectively. T2* represents the entire subclade T2

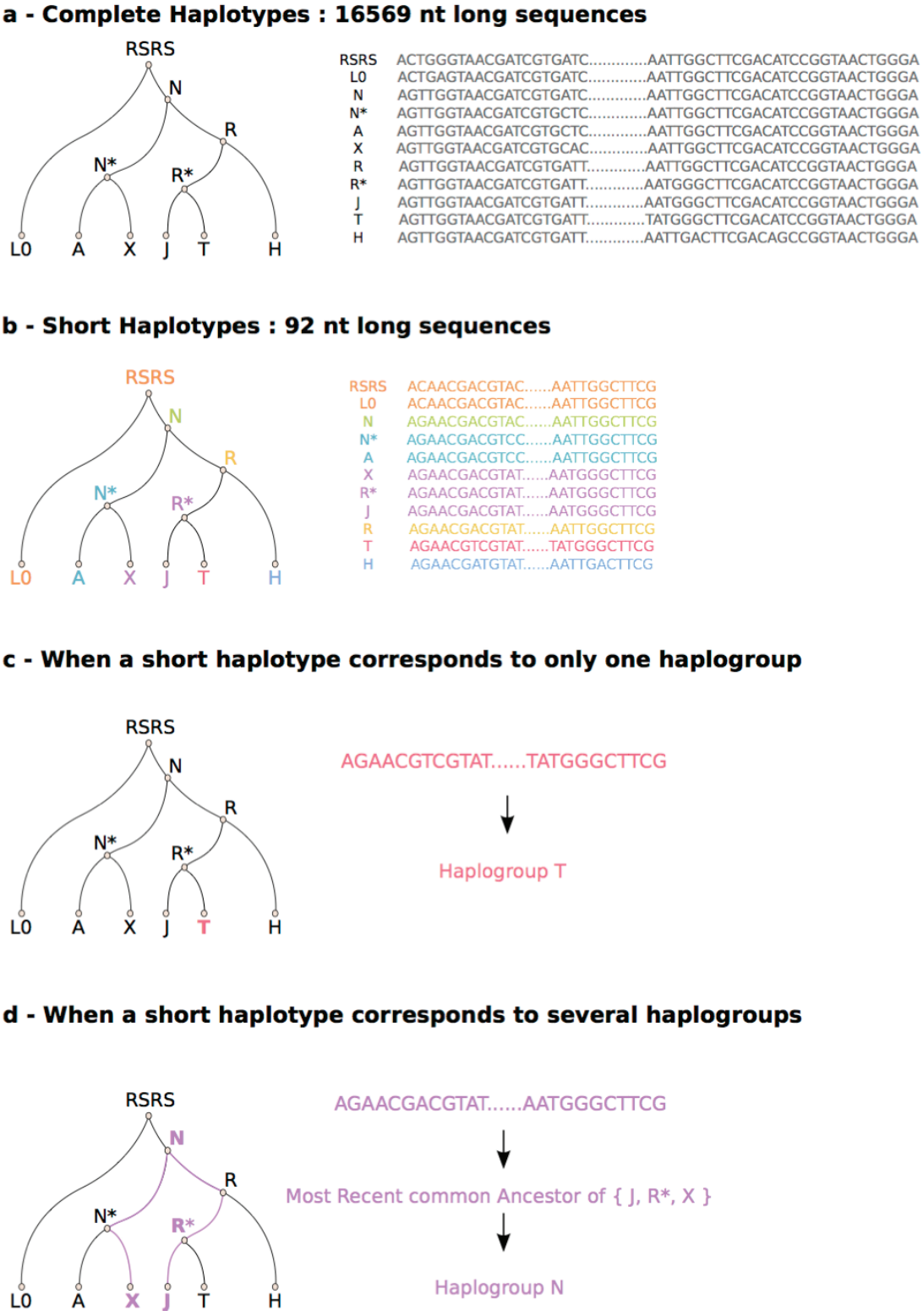


Figure 1: Simplified representation of the phylogenic method used to infer haplogroups.
a. Full-length haplotypic sequences are reconstructed at each node of the reference tree. **b.** Haplotypes are then restricted to available loci. Sequences of the same color are identical. **c.** Unique short haplotypes are matched directly with the corresponding haplogroup. **d.** Sequences matching with several haplogroups are associated to their Most Recent common Ancestor haplogroup.

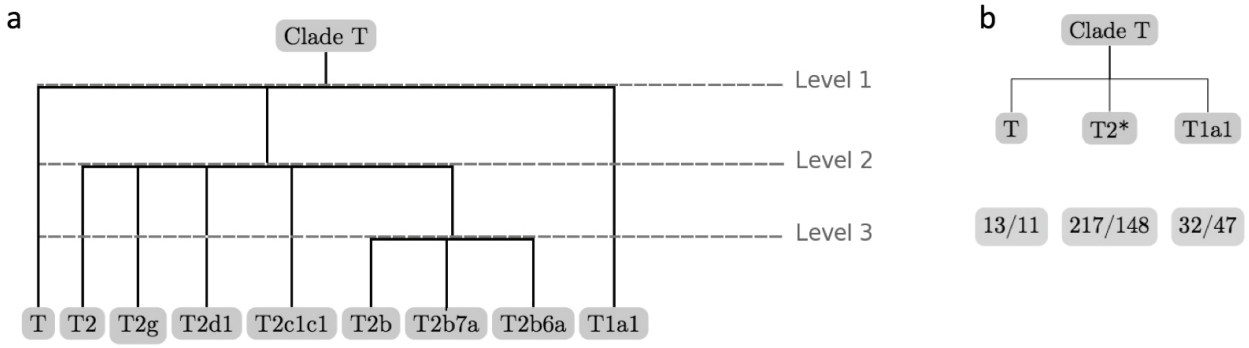


Figure 2: Phylogenetic tree of subclade T tested for association with ALTree
a. Phylogenetic tree of subclade T with all observed haplogroups. A homogeneity test is performed at each level of the tree.
b. First level of phylogenetic tree of subclade T. Averaged counts from resampling of affected and unaffected are indicated below each subclade, respectively. T2* represents the entire subclade T2.